Conseil national
de l'information statistique

Paris, le 12 juillet 2022 – n° 100/H030

PANELS ET COHORTES STATISTIQUES, QUELS NOUVEAUX OUTILS POUR ECLAIRER LE DEBAT PUBLIC ?

_____

Compte rendu du colloque du 18 mai 2022

COMPTE RENDU DU COLLOQUE

PANELS ET COHORTES STATISTIQUES,
QUELS NOUVEAUX OUTILS POUR ECLAIRER LE DEBAT PUBLIC ?

- 18 mai 2022 -

Président : Patrice DURAN

RAPPEL DU PROGRAMME

OUVERTURE 7

INTRODUCTION 8

I. L'ECHANTILLON DEMOGRAPHIQUE PERMANENT, UNE SOURCE MULTITHEMATIQUE 10

II. SESSION THEMATIQUE 1 : JUSTICE ET SECURITE 12

1. Le panel des jeunes suivis en justice 13

2. Projet PARCOURS, refonte du système d'information de la DPJJ 13

3. Les enjeux statistiques de la généralisation des procédures pénales numériques 14

III. SESSION THEMATIQUE 2 : ENTREPRISES 16

1. Réseau CompNet 17

1. L'utilisation des panels en statistiques d'entreprises 18

IV. SESSION THEMATIQUE 3 : EDUCATION ET FORMATION 20

1. Panels et cohortes statistiques dans le domaine de l'éducation 20

2. Panels et cohortes pour éclairer l'action publique en matière de scolarisation des jeunes en situation de handicap 21

V. TABLE RONDE : LES PANELS DANS LE DOMAINE DE LA SANTE ET LEUR GOUVERNANCE 24

CLOTURE 31

Liste des participants

ADAM	Lorraine	PROGEDO
ALAC	Rojda	Fédération nationale solidarité Femmes
ALLAIN	Samuel	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
BAGEIN	Guillaume	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
BARHOUMI	Meriam	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
BAUCHAT	Barbara	Ministère de la Culture - Département des études, de la prospective, des statistiques et de la documentation (DEPS-Doc)
BELLOC	Brigitte	Société française de statistique (SFdS)
BERTHOMIER	Nathalie	Ministère de la Culture
BODIN	Jean-Louis	CESD - Statisticiens pour le Développement
BONDON	Marine	Institut national des études démographiques (INED)
BONNEVIALLE	Lionel	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques (Sies)
BOUARFA	Naima	Mairie de Gennevilliers
CAILLAUD	Alain	Particulier
CARON	Nathalie	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
CHALEIX	Mylène	Institut National de la statistique et des études économiques (INSEE) - Direction de la méthodologie et de la coordination statistique et internationale (DMCSI)
CHARRANCE	Géraldine	Institut national des études démographiques (INED)
CHEVALIER	Pascal	Ministère de la Justice - Sous-direction de la statistique et des études
COLIN	Christel	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
COLIN	Paul	Centre national de la recherche scientifique (CNRS)
D'ALESSANDRO	Cristina	Institut National de la statistique et des études économiques (INSEE) – Direction de la diffusion et de l'action régionale (DDAR)
DEFRESNE	Marion	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
DELAUNAY	Isabelle	Conseil départemental du Vaucluse
DELTA	Lionel	Institut national de la statistique et des études économiques (INSEE) - Division des sondages
DI MAURO	FILIPPO	Réseau CompNet / University of Singapore
DIACAR	Vanessa	Fédération nationale solidarité Femmes
DJELLAL	Faridah	Université de Lille
DUBOIS	Marie-Michèle	Conseil national de l'information statistique (CNIS)
DURAN	Patrice	Ecole normale supérieure
ELBAUM	Mireille	Ministère des Solidarités et de la santé - Inspection générale des affaires sociales (IGAS)
FERRY	Odile	Observatoire national de la vie étudiante (OVE)

FOURRE	Marie	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques (Sies)
FRESSON-MARTINEZ	Catherine	Ministère de l'Agriculture et de l'alimentation - Service de la statistique et de la prospective (SSP)
GOLDBERG	Marcel	Institut national de la santé et de la recherche médicale (INSERM)
GONZALEZ-DEMICHEL	Christine	Ministère de l'Intérieur - Service statistique ministériel de la sécurité intérieure (SSMSI)
GUILLAUMAT-TAILLIET	François	Conseil national de l'information statistique (CNIS)
GUILLEMOT	Danièle	Institut National de la statistique et des études économiques (INSEE) – Direction des statistiques d'entreprises (DSE)
GURGAND	Marc	École d'économie de Paris
HAAG	Olivier	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
HADDAK	Mohamed Mouloud	Université Gustave Eiffel
HADJ LARBI	Alisée	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
HENRY	Jade	Ministère de l'Intérieur - Département des statistiques, des études et de la documentation (DSED)
HUET	Thomas	Institut national des études démographiques (INED)
IGUERTSIRA	Hayet	Mairie de Paris
JOURDAN	Virginie	Ministère de l'Intérieur - Département des statistiques, des études et de la documentation (DSED)
KLIPFEL	Justine	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques (Sies)
LAURIEUX	Patrick	Particulier
LE MINEZ	Sylvie	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
LECOUVEY	François	Centre d'études et de recherches économiques sur l'énergie (CEREN)
LIXI	Clotilde	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques (Sies)
MANSOURI GUILANI	Nasser	Confédération générale du travail (CGT)
MAREAU	Quentin	Conseil départemental de Meurthe-et-Moselle
MARKOU	Efi	Institut national des études démographiques (INED)
MAROUTEIX	Olivier	Conseil départemental de l'Essonne
MARTIN	Hugues	Ministère de la Justice
MAUREL	Françoise	Conseil national de l'information statistique (CNIS)
MONTUS	Arnaud	Conseil national de l'information statistique (CNIS)
MOREAU	Sylvain	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques d'entreprises (DSE)
MOUNIER	Lise	Centre Maurice Halbwachs CNRS
MÚNERA LÓPEZ	Manuela	France Volontaires

NAIT IGHIL	Lyes	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Sous-direction des systèmes d'information et des études statistiques (Sies)
NINNIN	Louis-Marie	Ministère de l'Intérieur - Département des statistiques, des études et de la documentation (DSED)
NUTARELLI	Mathilde	Centre d'études de l'emploi et du travail (CEET)
ORLANDI	Jean-Yves	Ministère de la Justice
OURLIAC	Benoît	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
PERRON	Zoé	Institut national des études démographiques (INED)
PETIT	Jean-Jacques	Particulier
PICARD	Sébastien	Ministère de la Culture - Département des études, de la prospective, des statistiques et de la documentation (DEPS-Doc)
POUILLARD	Denys	Observatoire de la vie politique et parlementaire
PRAT	Isabelle	Ministère de l'Intérieur - Service statistique ministériel de la sécurité intérieure (SSMSI)
PROKOVAS	Nicolas	Confédération générale du travail (CGT)
PUIG	José	Délégation interministérielle autisme et neurodéveloppement
RAFFAELLI	Christelle	Ministère de l'Education nationale, de la jeunesse et des sports
RAGHAVAN	Nicolas	Conseil départemental de l'Essonne
REY	Grégoire	Institut national de la santé et de la recherche médicale (INSERM)
REYNAUD	Bérengère	Ubiqus
RIMBEAULT	Chloé	Initiative France
ROBERT-BOBÉE	Isabelle	Institut National de la statistique et des études économiques (INSEE) - Direction des statistiques démographiques et sociales (DSDS)
ROCHER	Thierry	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
ROCHER	Guillaume	Ministère de l'Education nationale, de la jeunesse et des sports - Direction de l'évaluation, de la prospective et de la performance (DEPP)
SANTOS	Aurélie	Institut national des études démographiques (INED)
SCHUHL	Pierrette	Ministère de l'Enseignement supérieur, de la recherche et de l'innovation - Direction générale de l'enseignement supérieur et de l'insertion professionnelle (DGESIP)
SERIEYX	Yvon	Union nationale des associations familiales (UNAF)
SPRUYT	Emilie	Centre national de la recherche scientifique (CNRS)
SUESSER	Jan Robert	Ligue des droits de l'homme
TAGNANI	Stéphane	Conseil national de l'information statistique (CNIS)
TAUBIN	Alexandra	Caisse d'allocations familiales de Paris
THAUVIN	Patricia	Institut national des études démographiques (INED)
TOULEMON	Laurent	Institut national des études démographiques (INED)
TROGNON	Alain	Particulier
VALLET	Louis-André	Centre national de la recherche scientifique (CNRS)
VIGNALE	Mélanie	Centre d'études et de recherches sur les qualifications (CEREQ)
VIVIER	Géraldine	Institut national des études démographiques (INED)

YANKAN	Leslie	Ministère des Solidarités et de la santé - Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)
ZINS	Marie	Université de Paris
ZOLOTOUKHINE	Erik	Centre national de la recherche scientifique (CNRS)

La séance est ouverte à 9 heures 40.

OUVERTURE

Patrice DURAN, Président du Cnis

Bonjour et bienvenue à tous. Je suis très heureux de vous accueillir aujourd'hui pour ce colloque initialement prévu en 2020, puis reprogrammé en 2021, qui peut finalement avoir lieu au centre de conférence Pierre Mendès-France de Bercy : la présence des intervenants et du public en salle est un signe clair que nous avons récupéré une forme de normalité. Il nous paraît en effet important au Cnis de renouer avec la concertation en présentiel, lorsque cela est possible et souhaitable, sans pour autant exclure de recourir à la vidéoconférence lorsque la situation l'exige. Nous faisons au mieux pour nous adapter aux besoins et profiter des avancées technologiques accélérées par la pandémie.

Après la rencontre du Cnis du 28 janvier dernier, consacrée aux appariements de données, aux questions d'ordre technique et méthodologique, mais aussi éthique et sociétal qu'ils soulèvent, nous sommes aujourd'hui réunis pour faire le point sur les avancées récentes sur les panels et les cohortes de la statistique publique et plus généralement sur les approches longitudinales. C'est là une suite logique. Les appariements permettent de traiter la transversalité des rapports des acteurs concernés par des problèmes publics débordant largement les nomenclatures politiques et administratives et dont le traitement nécessite de fait une coordination appropriée. Sans nul doute, les appariements sont vite apparus comme des instruments précieux d'une telle coordination, mais il convenait maintenant de réintroduire l'intérêt et l'importance de la dimension temporelle dans la production de données afin d'en préciser la portée comme la qualité. Comment penser l'articulation du passé, du présent et de l'avenir sans une contextualisation historique précise ? En quoi ces approches longitudinales répondent-elles aux besoins de connaissance ? Quelles sont les principales évolutions des systèmes d'information ? Peut-on s'inspirer d'expériences menées à l'étranger ? … Et mieux nous comparer à nos partenaires ?

Les panels et cohortes sont de fait des outils précieux et même incontournables pour appréhender les trajectoires des individus et des groupes au fil du temps. L'observation des parcours de vie dans le but d'appréhender les enjeux sociétaux, quelle que soit la problématique ou la thématique abordée, est par ailleurs une préoccupation constante au Cnis depuis de nombreuses années. Je ne m'attarderai pas sur la vision rétrospective des approches longitudinales de la statistique publique depuis les premières expériences pionnières des années 1970. Un numéro de Chroniques du Cnis, publiée en février 2021, retrace à grands traits cette évolution jusqu'en 2020. On peut cependant en évoquer rapidement les principales dimensions.

- Bien que le moyen terme actuel du Cnis (2019-2023) ne comporte pas d'avis spécifique sur ces outils, des avis généraux et des avis de commissions thématiques (Démographie, Services publics et Emploi par exemple) recommandent le recours à des approches longitudinales pour une mesure plus approfondie des problématiques sociales.

- En suivant des individus et en étudiant les transitions individuelles, on peut quantifier des stocks mais aussi les flux, bref mieux comprendre et interpréter des phénomènes complexes, comme les parcours dans la grande pauvreté, les décrochages scolaires, l'entrée dans la vie active, la sortie du chômage, les parcours des jeunes en prise avec la justice, les entrées dans la dépendance.

- La santé, l'éducation, l'emploi et la formation sont les thèmes qui rassemblent le plus de panels actuellement recensés dans le champ du Cnis.

- Un des domaines au centre de l'attention et des attentes des utilisateurs et dans lequel des progrès sont attendus est notamment celui de l'usage de sources administratives multiples pour constituer des panels ou de leur appariement avec des panels existants comme l'EDP.

- Il existe à ce jour peu de panels ayant une dimension internationale comme l'enquête Statistiques sur les Ressources et Conditions de Vie (SRCV), l'enquête sur la Santé, le vieillissement et la retraite en Europe (Share) ou l'enquête Emploi. Les organisations internationales (comme l'OCDE ou la Banque mondiale) pourront-elles avoir un rôle plus important dans le financement des expériences internationales, ainsi que dans la gestion d'outils de plus en plus ambitieux ? Comment le cadre juridique devra-t-il évoluer pour

développer l'utilisation de ces données dans le respect de la confidentialité des données ? Ce sont là quelques-unes des questions qui seront débattues aujourd'hui.

Aujourd'hui les enjeux de connaissance sont devenus déterminants pour la maîtrise et le pilotage de l'action publique. Et il convient de ne pas oublier que l'Insee, à la différence de ses homologues d'autres pays, n'est pas seulement un institut de statistique, il est aussi un producteur d'études économiques, et on pourrait dire aujourd'hui plus largement de sciences sociales. Et ceci est également vrai pour les services statistiques ministériels. Il convient du même coup de ne pas avoir une vision étroite des questions méthodologiques. Celles-ci sont toujours à rapporter aux questions posées et modalités de problématisation à travers lesquelles elles ont été définies. En retour, il est bien clair que les conclusions des études produites sont inséparables de la démarche qui a permis de les établir. Elles constituent la preuve de sa fécondité et aussi de ses limites, voire de ses insuffisances.

L'introduction de la démarche temporelle n'est pas sans difficultés. Ne serait-ce que parce qu'elle vient rappeler l'historicisation des catégories et des enquêtes statistiques tout comme leur appartenance à des contextes institutionnels spécifiques. Ce qui nous conduit là encore à nous interroger sur les conventions d'équivalence du fait de façons de classer différentes tant au plan international qu'à celui de l'histoire. Il convient donc d'être prudent en particulier avec l'ouverture de plus en plus grande aux sources administratives qui sont essentielles pour les panels en particulier, mais posent en effet la question de leur cadrage institutionnel d'origine qui ne facilite guère la comparabilité entre pays, en Europe tout particulièrement. De plus, si les sources administratives sont déterminantes pour les panels. Elles renvoient aussi à des questions nouvelles, le secret, la sécurité des personnes, l'anonymisation. C'est bien pour cela que le code statistique non signifiant peut présenter de grandes vertus et constitue aujourd'hui un atout indiscutable.

Si la réflexivité est une dimension de plus en plus significative de l'action publique, on ne peut oublier que la maîtrise intellectuelle du monde ne nous donne pas pour autant sa maîtrise pratique qui, elle, est affaire d'action collective. Notre capacité d'intervention ne peut en effet se réduire à la seule capacité à rendre le monde intelligible. Aussi, le lien entre connaissance et action conclura justement cette journée à travers une table ronde sur les panels dans le domaine de la santé qui abordera les questions de gouvernance et de coordination en prenant appui sur un secteur, celui de la santé, qui s'est considérablement ouvert sur de multiples dimensions qui induisent des modalités de coopération tout à la fois élargies et peu réductibles à un simple traitement hiérarchique.

Je vous remercie encore pour votre participation et pour les échanges dont je ne doute pas qu'ils seront fructueux. J'en profite pour remercier les organisateurs d'avoir œuvré et su faire preuve de patience pour que ce colloque puisse aujourd'hui se dérouler dans les meilleures conditions. Je vous souhaite une très bonne journée et j'espère que vous l'apprécierez autant que la précédente rencontre sur les appariements.

INTRODUCTION

Laurent TOULEMON, Institut national d'études démographiques

Bonjour à tous. Je remercie les organisateurs de me donner l'opportunité d'introduire cette journée qui se veut complémentaire de celle des appariements.

Le rapport rédigé en 2004 par Mylène Chaleix et Stéfan Lollivier se voulait volontariste. De plus en plus de données administratives étaient disponibles. Ce rapport a permis à l'Insee de s'emparer de ce changement pour améliorer de façon considérable les statistiques. Le premier point portait sur la complémentarité entre enquêtes et données administratives. Nous l'avions vu avec l'enquête Emploi. Des tentatives avaient déjà eu lieu pour obtenir plus facilement qu'en posant la question des données précises sur les revenus, à partir des données fiscales.

Ensuite, le rapport portait l'idée de réaliser des panels sur tous les âges, avec trois groupes d'intérêt : les enfants, les adultes, les personnes âgées. Les trois panels ont été réalisés : Elfe pour les enfants avec la cohorte 2011, SHARE pour les plus de 50 ans, et surtout l'échantillon démographique permanent (EDP) qui s'est taillé la part du lion, avec un quadruplement de l'échantillon et l'enrichissement avec les données administratives pour introduire les données fiscales. La conclusion du rapport présentait aussi deux questions importantes : le respect de la vie privée et la collecte loyale (prévenir les répondants de l'information recueillie sur eux), ainsi que la collaboration avec les chercheurs. L'Insee est une direction du

ministère de l'économie et de finances, avec des corps professionnels, c'est une force mais il est moins ouvert sur la recherche que dans d'autres pays. Il apparaissait alors utile pour l'Insee de s'ouvrir vers le monde académique en France comme à l'étranger et vers les autres instituts de statistique.

De plus en plus de données sont disponibles et des progrès considérables ont été réalisés en termes de panels. Les panels ont différentes utilisations. Ils servent d'abord à recueillir des données longitudinales pour avoir une vision de l'évolution des cohortes. Au-delà de l'hétérogénéité entre les individus, le panel introduit l'hétérogénéité de l'individu lui-même dans le temps. Les individus sont comme ils sont sans qu'on puisse vraiment l'expliquer. Cependant, on va mettre en regard les événements qu'ils subissent avec l'idée que le passé explique le futur. Cela permet de proposer une inférence causale. Il existe une ambition très forte d'arriver à un schéma causal, mais l'on oublie l'hétérogénéité des individus. Cela amène à des études très intéressantes, mais qui ne concernent personne. Les panels sont beaucoup plus que cela. Le fait de compléter les caractéristiques instantanées permet de se poser de nouvelles questions. Prenons l'exemple des inégalités : se renforcent-elles les unes les autres au cours du temps, ou bien observe-t-on des effets de rappel ? Les politiques sociales permettent-elles de rattraper les personnes qui passent pas une phase de fragilité ou de pauvreté ? À quel moment est-il le plus efficace d'agir sur les inégalités ? Cela fait aussi le lien avec les questions d'équité, d'égalité des chances, d'héritabilité des chances.

Les panels offrent ainsi des résultats nouveaux et importants, où les évolutions temporelles sont au cœur de la définition des situations et de la réflexion sur les politiques envisageables. Citons ici quelques exemples. Sur les situations difficiles, la question de savoir si ce sont des situations stables exige un suivi des personnes dans le temps. L'étude récente de Aliocha Accardo a montré que les personnes restent largement dans leur position initiale en termes de niveau de vie. Les données fiscales de l'EDP montrent en effet moins de changements sur 9 ans que ce que l'on pourrait déduire des changements d'une année sur l'autre, et qu'il existe donc un effet de rappel qui limite les mobilités sur le long terme. C'est moins vrai dans l'enquête SRCV qui fait apparaître de nombreux changements. Les travaux de l'Insee ont été très riches en s'intéressant à de nombreuses dimensions de la pauvreté et à leur évolution au cours du temps.

Sur les personnes sans domicile fixe, l'approche longitudinale permet également de réfléchir aux politiques possibles. Les politiques doivent-elles être menées en amont, pour éviter que les personnes ne se trouvent en grande difficulté face au logement ? Lorsque les personnes sont en grande difficulté et déjà à la rue ? Faut-il mettre l'accent sur la sortie de la situation et l'hébergement pérenne des personnes sans domicile ? Les politiques sont différentes, les acteurs aussi. De même, faut-il éviter que les personnes aillent en prison ? Faut-il préparer la sortie et la réinsertion ? En termes de récidive, l'évolution des personnes à l'issue de leur séjour en prison va permettre de juger de l'efficacité de la prison.

Sur l'espérance de vie, on voit grâce aux travaux de Nathalie Blanpain à partir de l'EDP, qui s'appuie sur les données fiscales et données de mortalité regroupées, que les pauvres meurent avant les riches. Ce constat est assez violent. Les différences ont un caractère un peu scandaleux et appellent une politique publique. L'idée que tout le monde est égal face à la santé en France est un mythe.

Les panels se heurtent à des contraintes techniques et institutionnelles spécifiques. Outre leur coût élevé, les panels par enquête souffrent d'attrition, des personnes étant perdues de vue entre les vagues d'interrogation. De ce point de vue, les données administratives sont extrêmement précieuses en permettant de renforcer les enquêtes, si les individus sont suivis au cours du temps. Grâce à elles, on va pouvoir faire un suivi passif sur le comportement des individus et on va conserver leurs données de contact. Les données administratives peuvent ainsi servir de support aux enquêtes en panel, mais elles posent la question du respect des personnes, qui ne fournissent pas leurs informations en acceptant qu'elles soient utilisées à des fins d'étude ou pour les interroger plus facilement. En pratique, ces appariements entre données administratives et enquêtes apparaissent compliqués. En témoigne la cohorte Elfe qui devait être enrichie avec les données de l'EDP ou du ministère de l'Éducation. Or ces démarches sont toujours au stade de projet.

Une autre difficulté de l'analyse des données de panel tient au fait que les personnes ne vivent pas tant d'événements que cela au cours de leur vie. Il faut donc un suivi dans le temps long ou avec un gros échantillon pour pouvoir tirer de vrais enseignements des données.

D'un point de vue institutionnel, des progrès gigantesques ont été accomplis, mais ils sont restés assez lents. Sur l'appariement de l'EDP avec les données de santé le projet, lancé à plusieurs reprises, devait prendre 3-4 ans. Aujourd'hui, un appariement a été fait par la Drees, mais dans des conditions telles que les données ne pouvaient pas être conservées au-delà de 5 ans. On voit ici une tension très brutale entre le

respect des personnes et l'ambition de construire des panels. Si l'on veut travailler sur les inégalités sociales de long terme, il faut pouvoir conserver des données personnelles sur le long terme, ce qui implique de gérer des risques en termes de respect de la vie privée et de droit à l'oubli. La distinction entre appariements à but statistique et gestion centralisée des fichiers administratifs devrait être mieux comprise. Il ne faut pas non plus négliger le problème politique qui se pose dans les données de santé, qui prennent une nouvelle dimension quand elles sont complétées par des données socio-économiques.

Concernant ces aspects institutionnels, le Cnis pourrait jouer un rôle moteur. L'Insee a fait un très gros effort de connaissance sur les pratiques à l'étranger et dans la recherche, mais les efforts doivent être poursuivis. De ce point de vue, le fait que le code statistique non signifiant (CSNS) n'ait pas été présenté avant d'être mis en place pour le seul périmètre de la statistique publique est un peu une occasion manquée. Pour moi, nous perdons 5 ans en termes d'accès pour la recherche. Le CASD permet, en respectant les contraintes du secret, dans des conditions parfaitement sécurisées, de donner accès à des données non anonymes pour les chercheurs, mais aussi pour des agents de l'Insee, s'agissant des données d'autres ministères. Nous pouvons espérer que le CASD pourra jouer le rôle de tiers de confiance sur les cohortes et panels. C'est un progrès immense et un outil très performant.

Nous aborderons aujourd'hui des thèmes très divers. En terminant cette introduction, je voudrais évoquer brièvement le projet LifeObs porté par l'Ined avec l'Insee et différentes universités pour rassembler trois panels européens, les enquêtes SHARE pour les personnes âgées, Generation and Gender Programme (GGP) pour les adultes et GUIDE-Eurocohort, deux nouvelles cohortes d'enfants homogènes à l'échelle européenne, recrutées l'une à la naissance, l'autre à 8 ans. Ces trois projets sont rassemblés dans un grand équipement structurel pour la recherche auquel l'Insee participe, avec l'enquête Familles de 2025 et des travaux méthodologiques sur la collecte multimode, l'appariement entre enquêtes et données administratives, et le suivi des personnes. Voilà un exemple de collaboration entre un institut de recherche, des universités et l'Insee.

I. L'ECHANTILLON DEMOGRAPHIQUE PERMANENT, UNE SOURCE MULTITHEMATIQUE

Isabelle ROBERT-BOBÉE, Insee

L'EDP (échantillon démographique permanent) est un échantillon d'individus qui existe depuis 1968 avec une finalité statistique pour l'étude des trajectoires, notamment des parcours résidentiels et des parcours sociaux. Ce panel est constitué de différentes sources de données pour donner un éclairage sur différentes dimensions pour un même individu. Multidimensionnelle, cette source est très riche en information, et elle constitue de ce fait plutôt une source « experts ».

L'EDP avait pour but de constituer un panel d'individus. Plusieurs approches sont possibles pour constituer des données de trajectoires : enquêter des personnes à un moment donné en leur posant des questions sur leur parcours passé (enquête rétrospective, faisant appel à la mémoire des enquêtés) ; interroger plusieurs fois les mêmes personnes pour recueillir des informations au fil des années (mais avec la difficulté de retrouver les personnes pour les réinterroger, et une attrition qui augmente donc au fil des années) ; ou, comme le panel EDP, rassembler au fil des années des données recueillies par ailleurs. Cette façon de constituer l'EDP a permis d'éviter le problème des biais de mémoire. Il permet aussi de combiner une diversité de source de données. Le dispositif est relativement économe, puisqu'il n'implique pas de coûts de collecte ni de cout de suivi.

A l'origine, nous avons retenu un critère très simple de sélection des individus. Appartiennent à l'EDP les individus nés les 4 premiers jours d'octobre, dits « jours EDP ». Dès la création du panel EDP, sont entrées des personnes nées à ces jours, quelle que soit l'année. Dès 1968, ce sont donc 500 000 personnes de tous âges qui ont intégré l'EDP. A l'origine, l'échantillon était composé de deux sources : les recensements de la population et l'état civil. L'entrée s'est faite massivement par le recensement de 1968 et quelques personnes sont entrées par l'état civil. L'EDP, dès l'origine, a été constitué non pas comme une cohorte, mais comme un ensemble d'individus sur des tranches d'âge très diversifiées. Sans attendre de nombreuses années, l'EDP pouvait donc déjà fournir des analyses statistiques intéressantes sur l'ensemble de la population.

Il est important qu'un panel puisse se renouveler, pour rester représentatif de l'ensemble de la population. L'EDP se renouvelle. Au fil du temps, on peut entrer dans l'EDP par un recensement, par des événements

d'état civil (naissances par exemple). Ce panel donne un suivi démographique des personnes. Par exemple, l'EDP permet de connaitre leur lieu de résidence, leur catégorie sociale et leurs diplômes à différentes dates. L'individu est suivi d'un point de vue statistique jusqu'à son décès ou son départ du territoire, les sources alimentant l'EDP étant des sources sur les personnes résidant en France.

Les sources qui alimentent l'EDP ont beaucoup évolué et l'EDP s'est adapté. Le recensement était une source exhaustive jusqu'en 1999, réalisé environ tous les 10 ans. Désormais, il repose sur des enquêtes annuelles qui chaque année, recense un échantillon d'adresses. Nous n'avons donc plus l'information en même temps pour tout le monde. Or il est important de disposer d'informations sociodémographiques pour tous les individus du panel. Deux adaptations ont donc été opérées pour s'adapter à cette évolution. Pour compenser la dégradation de la qualité des estimations du fait de la réduction des données disponibles une même année, la taille de l'échantillon a été augmenté : le nombre de « jours EDP » est passé de 4 à 16. Nous nous sommes aussi tournés vers deux sources de données : les panels « tous salariés » et « tous actifs » et les données socio-fiscales.

Les données socio-fiscales sont exhaustives et annuelles. Cela a enrichit l'EDP de deux manière. Au-delà des revenus et des niveaux de vie, ces données offrent notamment des informations sur la situation familiale des personnes, le calcul de l'impôt faisant intervenir cette dimension. Elles permettent aussi d'enrichir l'EDP par une variable socio-économique importante, le niveau de vie. Les différentiels sociaux abordés souvent en termes de catégorie sociale et de de diplôme, peuvent désormais l'être aussi dans l'EDP selon le niveau de vie. Ces trois dimensions ne sont pas résumables en une seule comme le montrent les nombreuses études réalisées. Une autre source a également été ajoutée dans l'EDP, le fichier électoral (inscription sur les listes électorales).

L'EDP est donc très riche en termes d'informations, à la fois dans la dimension temporelle et par la richesse des différentes sources qui l'alimentent, qu'il s'agisse des sources historiques ou plus récemment par exemple des niveaux de vie. De ce fait, exploiter les données de l'EDP n'est pas simple. C'est la contrepartie de la richesse des informations. L'EDP va nécessiter, pour le chercheur ou le chargé d'études, d'être au clair sur sa population d'intérêt. Il faut prendre des précautions pour s'assurer que la population d'intérêt n'a pas été biaisée lors de sa constitution par les différentes sélections de population qu'il va opérer dans les différentes sources de l'EDP (personnes présentes dans telle source avec tel critère de sélection, et dans telle autre avec tel autre critère de sélection etc.) pour réaliser son étude statistique. Il faut aussi se poser des questions sur le moment auquel on a besoin pour l'étude d'informations sur les personnes. Par exemple, du fait de la méthode et de la modalité de collecte du recensement, les informations ne sont pas disponibles au même moment pour tous les individus EDP. Une année donnée, environ 1/7ème des individus du panel EDP est interrogé à une enquête de recensement donnée. Peut-on aller chercher l'information recueillies à une enquête annuelle passée pour élargir la taille de la population d'intérêt ? faut-il aller chercher l'information dans une autre source (données fiscales par exemple) ? Il faut se poser ce type de question quand on exploite les données de l'EDP. Il y a donc une réflexion préalable plus poussée pour exploiter l'EDP que pour l'exploitation de données d'enquêtes statistiques « traditionnelles », mono-source et en coupe. C'est pour cela que je qualifiais en introduction l'EDP de source pour un usage « expert ».

Par exemple, l'Insee a réalisé une étude sur l'évolution du niveau de vie lors du passage à la retraite. Il n'existe pas dans l'EDP de variable permettant d'identifier un passage à la retraite. Il faut donc constituer un indicateur. Nous l'avons fait à partir de montants de pension déclarés dans les données socio-fiscales. Or ce sont des données annuelles. Retenons-nous comme définition de l'année de départ à la retraite l'année où un individu perçoit ne serait-ce que 1 € de pension ? Quel seuil retenir sinon ? Nous avons réalisé différents tests et comparé avec les flux de retraités bien connus de la Drees. Il faut mener une démarche de construction et de comparaison. Une autre complexité nouvelle avec le passage du recensement exhaustif à des enquêtes annuelles de recensement sur échantillon est l'usage des pondérations : elles ont été introduites dans l'EDP, chaque individu ne représentant plus le même nombre d'habitants et il faut en tenir compte dans l'exploitation de l'EDP.

Les données pour la recherche sont accessibles de manière spécifique via le CASD après une démarche auprès du comité du secret. La source est complexe, mais depuis l'ouverture des données par le CASD, nous avons vu une augmentation de l'accès et des projets utilisant ces données. Actuellement, 60 projets de recherche incluent les données de l'EDP. Ces données étant difficiles à traiter, l'Insee a mis en place un groupe d'utilisateurs pour un partage des pratiques et aider à l'exploitation des données.

Aujourd'hui paraît une nouvelle publication exploitant les données de l'EDP sur une thématique nouvelle, étudiant la façon dont les niveaux de vie évoluent entre les parents et les enfants, étude inédite, possible

pour la première fois grâce à ce panel et qui montre une nouvelle fois que l'inclusion du niveau de vie dans l'EDP a ouvert de nouveaux champs d'analyses statistiques des trajectoires.

Yvon SERIEYX, UNAF

Au vu du processus de constitution des échantillons, serait-il techniquement possible, dans le cadre d'un projet de recherche, d'effectuer une reconstitution *ad hoc* de données ? Si l'événement d'état civil était la 1ère naissance, serait-il possible de constituer un échantillon de dyades qui remonterait pour les deux membres du couple en amont ? Des personnes célibataires se sont-elles déjà mises en couple avec une autre personne du panel ?

Isabelle ROBERT-BOBEE

Le suivi porte sur les individus nés les « jours EDP ». Par exemple, si deux frères jumeaux sont dans l'EDP, les deux sont suivis dans le panel. Si deux personnes en couple sont nées lors des jours EDP, elles sont également suivies. Le nombre de telles « paires » a fortement augmenté en passant à 16 jours de naissance. Nous commençons à pouvoir coupler et suivre les trajectoires.

L'EDP intègre des informations caractérisant les personnes nées des jours EDP, notamment des données sociodémographiques concernant des personnes qui résident avec elles (informations du recensement par exemple). Si ces personnes ne résident ensuite plus elle elles, elles ne sont pas suivies dans le panel. Mais des études sont possibles. Par exemple, nous avons mené à l'Insee une étude sur l'évolution du niveau de vie après une séparation. Nous avions l'information sur des femmes et des hommes nés des jours EDP qui étaient en couple à un moment donné et avons étudié le niveau de vie de ces personnes quand elles étaient en couple et après leur séparation. Cela a permis d'étudier les évolutions séparément pour les femmes et pour les hommes, et d'apporter un éclairage statistique important à ce sujet, même s'il n'était pas possible de suivre au sein de chaque couple l'évolution du niveau de vie de chacun des conjoints, les deux n'étant pas forcément nés des jours EDP.

Laurent TOULEMON

Nous avons des panels d'individus. Les individus ont une identité, mais nous pouvons envisager aussi des panels de couples, de ménages. Nous le voyons dans les enquêtes SRCV. Si le couple se sépare, les deux personnes se séparent. Nous aurons les deux ménages. Quand on suit les couples, on peut vouloir suivre les deux membres du couple pour déterminer qui gagne/perd après la séparation. L'identité ménage ou couple, au moment de la rupture, au-delà de l'individu pose des questions relativement compliquées pour l'observation statistique. Avec des fichiers administratifs exhaustifs, nous avons tout le monde. Dans les enquêtes, si nous devions prendre tout le halo, nous aurions un échantillon un peu bizarre. Ainsi, les personnes avec beaucoup de ruptures sont beaucoup plus présentes dans l'enquête SILC et l'échantillon est assez fortement biaisé. Les problèmes de représentativité sont relativement compliqués.

Zoé PERRON, Ined

Les individus appartenant au panel sont-ils informés de leur appartenance à ce panel et des informations collectées ?

Isabelle ROBERT-BOBEE

Nous sommes en conformité avec le RGPD. Compte tenu du fait qu'il s'agit ici de l'utilisation de données déjà collectées, l'Insee diffuse une information collective. Les informations sont accessibles sur le site de l'Insee, qui précise les traitements statistiques de données personnelles mis en œuvre par l'institut ainsi que les droits associés.

II. SESSION THEMATIQUE 1 : JUSTICE ET SECURITE

Pascal CHEVALIER, Service statistique du ministère de la Justice

Nous avons choisi de vous présenter plusieurs dispositifs innovants. J'aurais pu citer les travaux du SSM menés en 2021 sur une cohorte de sortants de prison en 2016 de manière à mesurer la récidive. Mais nous avons convenu de vous présenter trois opérations innovantes avec de forts enjeux pour les années à venir :

le panel des jeunes suivis en justice, en lien avec un avis de moyen terme du Cnis sur les jeunes en prise avec la justice, la refonte du système d'information de la PJJ pour un suivi centralisé et dématérialisé des jeunes, répondant lui aussi à un avis de moyen terme sur l'avenir des enfants suivis par la protection de l'enfance et enfin, un dispositif autour de la mise en place de la procédure pénale numérique qui vise un suivi dématérialisé des pièces et surtout la mise en place d'un identifiant unique permettant un suivi longitudinal des affaires.

1. Le panel des jeunes suivis en justice

Le projet de panel des jeunes s'appuie sur le panel des mineurs qui existait jusque-là, décidé en 1996, mis en production en 2008 afin de décrire des trajectoires sociales et judiciaires des mineurs délinquants et des jeunes en danger. Ce panel a pu se développer jusqu'en 2013 en s'appuyant sur des rapprochements de données. A partir de 2013, ce panel a un peu périclité. Les applications de suivi des affaires pénales ont été supprimées pour être remplacées par le logiciel Cassiopée. Ce logiciel couvre ainsi la dimension pénale des affaires et Wineurs le civil.

Nous avions donc deux applications, Wineurs et Cassiopée, qui ne communiquaient pas entre elles et n'ont pas pu être rapprochées. Dès lors que l'accès aux données nominatives a été accepté pour le SSM Justice au travers de la loi de programmation de la justice et la révision de l'article 38-1 du Code de procédure pénale, la possibilité de réaliser des appariements entre les deux logiciels a été ouverte en vue de la mise en place d'un panel.

Cela va conduire à la mise en place de nouveaux flux de données. Aujourd'hui, le SSM a déjà accès à Cassiopée, mais ce flux complémentaire est essentiel car il va permettre l'accès aux données nominatives et à des données complémentaires pour avoir de l'information sur les trajectoires sociales des individus. Ce flux est en cours d'expérimentation.

Au-delà de ces problèmes pratiques, la relance de ce panel avec la LPJ permet d'envisager de nouvelles perspectives. Il a ainsi été imaginé un panel plus large que le panel des mineurs qui existait jusqu'alors, au travers d'un échantillon plus important, mais également avec une fenêtre de suivi plus large. Au-delà des individus nés entre le 1er et le 15 octobre, le choix des dates de naissance des individus sélectionnés a été élargi de manière à se caler sur l'extension de l'échantillon de l'EDP en vue de rapprochements futurs ; les individus nés les 4 premiers jours des trois trimestres restants ont ainsi été ajoutés. En termes d'âge, le choix d'élargir le panel aux jeunes adultes de moins de 26 ans impliqués dans une affaire pénale et les jeunes suivis en assistance éducative par le juge des enfants a également été retenu.

Dès que le flux de données sera opérationnel, le panel pourra être mis en place par appariement des fichiers de l'assistance éducative des mineurs et des affaires pénales. Ce panel présente un certain nombre d'avantages et offre des perspectives en termes d'études, d'analyses et de politiques publiques pour obtenir une vision plus large des trajectoires pénales (procédures alternatives, classements qui ne figurent pas dans le casier judiciaire). Il offrira aussi une plus grande richesse de suivi temporel. Le panel permettra aussi de répondre à des questions sur le lien entre enfance en danger et enfance délinquante. A cela s'ajoutent des perspectives de rapprochement avec les victimes, car ces jeunes sont aussi victimes d'infraction. Nous pourrions aussi envisager un rapprochement avec le système d'information de la PJJ, le dispositif PARCOURS pour appréhender les mesures prises dans ce cadre et avoir des informations plus larges. Enfin, à plus long terme, nous prévoyons des rapprochements avec d'autres sources sociodémographiques, notamment les sources fiscales et l'EDP.

2. Projet PARCOURS, refonte du système d'information de la DPJJ

Jean-Yves ORLANDI, ministère de la Justice

Le système d'information avait pour objectif initial de compter le nombre de jeunes et les prises en charge au sein des établissements et services de la Protection judiciaire de la jeunesse, en termes de décisions judiciaires, d'activités de jour (dispositif d'insertion de la DPJJ) et de suivis éducatifs en détention. A l'origine, il s'agit d'un fichier Excel. Il a été repris au niveau national et a fait l'objet de plusieurs versions, devenu GAME, GAME2000 puis GAME2010, passage d'une version monoposte à une version réseau pour partager les données en profitant du développement numérique. La PJJ s'est dotée d'un outil de plus en plus performant. Pour autant, l'outil avait encore besoin d'être amélioré, notamment techniquement pour gagner en capacité d'évolution. Le projet PARCOURS s'est donc révélé important, avec de nombreuses adaptations fonctionnelles. Comme pour GAME, toutes les données sont saisies au niveau des services, et

des unités parfois d'une dizaine d'agents. Toutes les données sont agglomérées dans un infocentre et en sont tirées, de manière mensuelle, sur l'intranet de la justice, des données sous forme de tableaux mis à jour mensuellement de toutes les informations concernant les jeunes (âge, sexe, décision, durées, échéances, etc.). L'exploitation est directe au niveau des directions interrégionales via un outil non ergonomique.

Le projet PARCOURS a été lancé en janvier 2019 avec trois objectifs assez marqués : refondre entièrement le système d'information pour une mise en conformité technique, fonctionnel et réglementaire, notamment au regard du RGPD, compte tenu de la sensibilité des données liées aux jeunes. La protection de l'enfant a été rénovée en 2016, et la justice des mineurs en matière pénale en 2021 avec entrée en vigueur du code la justice pénale des mineurs (CJPM). Par ailleurs, GAME avait 4 ans de retard, la dernière évolution étant intervenue en 2016. Il a fallu également prendre en compte l'ensemble des besoins des utilisateurs et viser l'objectif de dématérialisation du dossier de suivi du jeune avec toutes les informations le concernant. Enfin, l'infocentre lui-même en cours de refonte sera beaucoup plus ergonomique et permettra aux services de faire des études plus adaptées à leur situation locale, à partir d'un outil plus performant qu'Excel.

Cet outil est au service de l'évaluation. Pour cela, il faut collecter et exploiter les données. Le dossier dématérialisé ne peut être complété que par ceux qui connaissent le mieux les jeunes. GAME ne rendait que peu service aux personnels : toutes les données saisies ne pouvaient qu'être imprimées, et non réutilisées dans un document de type traitement de texte par exemple. PARCOURS vise à constituer un dossier partagé entre les différents acteurs, à l'avenir potentiellement ouvert aux agents de l'administration pénitentiaire, greffiers et magistrats dans l'échange d'informations et la transmission des rapports à fournir à l'instance judiciaire. Ces données permettent d'envisager une évaluation plus individuelle pour le jeune au regard du socle commun de compétences et de connaissances sur la base du décret de 2015. Enfin, il s'agit d'évaluer les interventions de la DPJJ dans les dispositifs de politiques publiques : la délinquance des mineurs et la prévention de la récidive, mais nous savons aussi qu'ils peuvent également être victimes et inscrits dans un dispositif de protection de l'enfance. Aujourd'hui, ces données sont encore difficiles à croiser.

Nous avons aussi un autre dispositif issu de la loi, la remontée de données vers l'ONPE pour étudier le parcours des jeunes, suivis concomitamment ou successivement par la PJJ et les Conseils départementaux.

L'application PARCOURS a été mise en service le 26 mai 2021 et il reste très probablement encore dix ans de travaux aujourd'hui pour atteindre l'ambition exprimée à l'origine du projet.

3. Les enjeux statistiques de la généralisation des procédures pénales numériques

Hugues MARTIN, ministère de la Justice

Il s'agit d'un programme interministériel. Nous avons dû démarrer par un chantier purement juridique qui a démarré en 2018. La 1ère année a été consacrée à la définition du dossier pénal numérique. Au-delà du dossier de procédure, la PPN vise à s'étendre jusqu'à l'exécution des peines. Nous travaillons avec la police, la gendarmerie et les juridictions, en particulier le parquet. La dernière loi de programmation justice avait un prisme nouveau sur le système d'information. La loi consacre, dans l'article 50, un dossier pénal numérique présentant la même valeur qu'un dossier papier. Ce texte nous a demandé énormément de travail avec notre responsable de la protection des données, les organes de la CNIL et du Conseil d'Etat puisque nous ouvrons un champ considérable de traitement des données. Le traitement peut potentiellement prendre toutes les données, de toute nature, qui peuvent permettre d'animer le travail du parquet et la décision judiciaire.

Le décret du 23 juin 2020 permet un traitement qui va supporter la procédure numérique et change le paradigme de la donnée. Les systèmes d'informations judiciaires portent actuellement des résumés de procédures (les données principales d'une procédure). Depuis une vingtaine d'années, ces données servent au SSM à avoir un minimum d'informations. Désormais, nous changeons de paradigme puisque nous visons l'exhaustivité de l'information en rendant numérique l'entièreté du dossier.

Pour des questions de cohabitation du papier et du numérique, nous n'avons pas choisi de formats originaux, restant sur des PDF. Nous faisons en sorte dès à présent, avec l'ensemble de nos partenaires, d'associer le format PDF A3 et la source du document, c'est-à-dire de la donnée intelligible d'un point de vue informatique pour avoir des données structurées directement. Il sera très difficile de prévoir des émissions structurées, mais nous pouvons au fur et à mesure avoir une sémantique pénale permettant

d'avoir une donnée de plus en plus compréhensible pour la statistique publique. Nous sommes dans les mêmes règles de confidentialité assez stricte, mais nous commençons à travailler avec la direction de programme et le SSM pour imaginer les travaux à venir.

Pour que ces pièces transitent et puissent être rapprochées à une procédure de justice, il nous faut un identifiant commun. Aujourd'hui, la procédure part physiquement au parquet. S'il manque un acte d'enquête, elle est physiquement retournée dans les services d'enquête. Ce programme devrait faciliter les transmissions. Dans le numérique, nous ne pouvons pas procéder de la sorte. Nous devons travailler sur un identifiant commun pour déterminer que nous sommes sur la même procédure. Cet identifiant est, à l'origine, purement opérationnel. Il présente une importance majeure.

Nous avons sorti l'identifiant de justice de la phase purement juridictionnelle. Nous l'avons externalisé, hébergé de manière externe et permettons à l'ensemble des clients potentiels, dont les services de police et gendarmerie, de demander un numéro. Cela permet surtout de créer un vrai référentiel sur les jonctions et disjonctions. L'outil existe, il reste à faire en sorte que l'ensemble des acteurs de la procédure pénale consomme ce référentiel commun.

A ce référentiel des affaires s'ajoute un méta-référentiel, dictionnaire sémantique pour agréger une réalité différente dans les dossiers. Les limites sont que la situation hybride papier/numérique perdurera longtemps. Il faudra également du temps pour obtenir une qualité de la donnée, liée à un exercice très important de caractérisation de la donnée, qui demande probablement une coordination européenne. Nous avons donc lancé un projet pour nous assurer que ce que nous construisons concorde avec l'émergence d'une interopérabilité des données au niveau européen.

Pascal CHEVALIER

Vous le voyez, ce projet est majeur.

Marie ZINS, Université Paris Descartes

Avez-vous d'autres exemples au niveau européen et international de regroupement de données ? Sur le projet PARCOURS, arrêtez-vous de suivre les mineurs dès lors qu'ils ont 18 ans ?

Hugues MARTIN

L'Italie lance un projet qui aboutira dans les mêmes jalons que les nôtres, travaillant avec les auxiliaires de justice. Les Espagnols s'y mettent aussi. La démarche en est à ses débuts. Nous commençons à pouvoir nous comparer entre pays de l'OCDE.

Jean-Yves ORLANDI

La PJJ ne prend en charge que les jeunes qui lui sont confiés par une autorité judiciaire. Sur le suivi civil, nous pouvons aller jusqu'à 21 ans. Côté pénal, certains dispositifs permettent à la PJJ de poursuivre le suivi au-delà de la majorité. Le jeune sera jugé par le tribunal pour enfants et accompagné par des personnels éducatifs.

Le code de justice pénale des mineurs est venu introduire une limite un peu plus floue sur cette majorité. Aujourd'hui, la mesure peut s'étendre jusqu'à 21 ans, voire au-delà. Ces mesures sont relativement rares. Il existe une grosse articulation entre PJJ et administration pénitentiaire.

Laurent TOULEMON

Le panel des mineurs suivis en justice sera-t-il alimenté par le suivi pénal ? Avez-vous aussi d'ores et déjà prévu un appariement avec l'EDP ? Faire des fichiers utilisables pour des études est un travail de construction. Cela fait-il déjà partie des réflexions ?

Pascal CHEVALIER

Nous avons du mal à fixer un calendrier pour la mise en place du panel compte tenu des difficultés pour mettre en place le flux d'informations et de données nominatives. Aujourd'hui, dans le meilleur des cas,

nous récupérerons les données fin 2022. Nous devrons étudier les appariements dans les prochains mois si ce flux se concrétise.

Notre priorité est de voir ce que nous pouvons récupérer du panel des mineurs qui existait jusqu'à présent. Coté affaires civiles, les remontées actuelles alimenteront le panel rétrospectivement pour constituer un panel historique si tout va bien, mais cela n'est pas garanti. Nous avons des bases gérées au niveau local. Chaque année, des juridictions ne parviennent pas à nous remonter des données en raison de difficultés principalement techniques. Nous essayons de voir dans quelle mesure nous pouvons mettre en place des remontées automatiques qui ne reposent pas uniquement sur le bon vouloir des juridictions. Nous sommes en phase d'expertise. Ce travail est prioritaire.

Quand ces questions seront résolues, nous envisagerons des appariements avec d'autres sources de données. Nous manquons quand même de visibilité sur la mise en place du panel et la récupération des données, mais l'enrichissement au cours du temps est important.

Mohamed Mouloud HADDAK, Université Gustave Eiffel

Je n'ai pas bien compris les hypothèses de travail, en particulier concernant la prévention. Quelles sont les hypothèses sur ces jeunes ? Les conditions sociales, l'échec scolaire, l'origine ethnique ? Vous parlez à la fois de mineurs et de jeunes. Quelle est la population témoin ? Par rapport aux indicateurs d'évaluation, que prévoyez-vous au-delà de la récidive ? Prévoyez-vous déjà des échanges ou des panels communs à l'échelle européenne, en particulier avec les pays qui ont travaillé sur ces questions comme la Suède ?

Jean-Yves ORLANDI

Sur les différents indicateurs, j'ai cité les trois principaux, qui font la une des journaux en période électorale. L'insertion sera un élément important. La loi récente qui pose l'obligation de formation entre 16 et 18 ans est l'un des axes sur lesquels nous allons travailler. Il s'agira de voir, dans les évaluations plus individuelles, combien de jeunes, à l'issue de notre prise en charge, sont mieux qu'avant celle-ci. Je n'ai pas cité les politiques au cœur du logement, un axe sur lequel la PJJ commence à être investie, de même que tous les aspects de santé, avec la difficulté que les données de santé sont plutôt sensibles. Il nous faut argumenter auprès de la CNIL pour travailler sur ces données. Avec l'ONPE, nous partageons des données pseudonymisées. Ces données ont pour ambition d'étudier les parcours des jeunes suivis successivement ou concomitamment par des dispositifs de suivi de la protection de l'enfance sur une durée de 20 ans. Aujourd'hui, la PJJ n'est pas encore en mesure de fournir ces données et seule la moitié des conseils départementaux arrive à remonter ces données.

Pascal CHEVALIER

Aujourd'hui, existait le panel des mineurs. Pour l'assistance éducative, nous avons continué à récupérer les données depuis 2008. Nous voudrions mettre en place un panel des jeunes qui s'appuie à la fois sur les données que nous avons pu collecter dans le cadre du panel des mineurs et qui prenne en compte les données pénales pour ces jeunes majeurs. Dès que nous pourrons apparier ces deux bases, nous aurons un premier panel prenant en compte toute la dimension historique collectée ces dernières années.

III. SESSION THEMATIQUE 2 : ENTREPRISES

Sylvain MOREAU, Insee

Après les individus, les ménages, nous allons parler des entreprises. Je vous présente les excuses de Faridah Djellal qui est souffrante. Filippo Di Mauro interviendra de Singapour.

La question des panels ne se posent pas en les mêmes termes que pour les ménages. Déjà, en France, nous avons la chance d'avoir le répertoire SIRENE et l'identifiant siren qui est utilisé par la totalité des administrations, ce qui facilite grandement le rapprochement de fichiers et quelque part permet de disposer sur un certain nombre de variables, notamment toutes les variables obtenues via les fichiers fiscaux, de panels « naturels ». Par ailleurs, toutes les entreprises ne se valent pas, suivant son secteur ou sa taille, certaines entreprises seront nécessairement regardées chaque année par la statistique publique. Puis, il existe des thématiques, pour lesquelles nous ne disposons pas de sources administrative, par exemple les questions d'innovation ou de développement durable, pour lesquelles nous avons besoin de mener des

études et la question de disposer de panel d'entreprises pour observer les changements de comportements se pose. Un exemple de pseudo panel : lors de la crise, l'INSEE a tiré un échantillon de 100 000 entreprises pour lesquelles il disposait des sources fiscales ce qui a permis d'avoir une vision très fine du choc que représentait la pandémie sur l'activité et même la trésorerie des entreprises, presque en temps réel. Mais avant d'essayer de répondre à toutes ces questions, il en est une fondamentale: qu'est-ce qu'une entreprise ?

Une entreprise est une combinaison d'entités légales qui jouit d'une autonomie de décision. Sur les individus, le nombre d'événements est relativement limité. Dans la vie des entreprises, des événements interviennent pratiquement tous les ans, ce qui implique un travail de consolidation extrêmement important et rend toute étude longitudinale extrêmement complexe.

Nous allons vous présenter quelques exemples pour vous montrer la façon dont nous travaillons et dont nous essayons d'élaborer un certain nombre d'outils pour représenter au mieux l'activité et la vie du système productif.

1. Réseau CompNet

Filippo DI MAURO, Réseau CompNet, University of Singapore

I am the chairman of Compnet, the competitiveness research network, an international network created more than 10 years ago. It does two things: first it provides a forum for productivity research and second, it has been creating what is now a relatively well known dataset of indicators for productivity analysis, based on firm level data, as Sylvain mentioned before. I should mention that INSEE is actually a member of Compnet and a very active one.

The motivation for what we have been doing during the last years is the fact that although micro data is available at the international level in Europe, it is relatively hard to access. This is not the case in France, but it is the case for several countries. We are now creating what we call the "micro data infrastructure" (MDI) within our Compnet network, which facilitates data access to this information, while respecting all the confidentiality requirements.

I will provide an example of an application for this data infrastructure to some rather detailed, granular data for the Netherlands. Researchers in the Netherlands have problems getting data access because they have to write a proposal, understand access procedures, pay costs and follow technical restrictions set by national statistical institutes (NSI). There are also many costs on the NSI side in allowing access to researchers, access procedures to be followed, dedicated systems that have to be provided to support researchers, as well as disclosure analysis to be put forward. Even when access to the data is possible, it is literally impossible to replicate such work in the international context because the legislation and data access vary from country to country.

Precisely what we want to do with this micro data infrastructure project is to facilitate the process. Our non-profit organisation is an intermediary between national statistical institutes and researchers to facilitate access to data. How does it work? On the one hand, national statistical institutes will provide access to run the code and the metadata, and authorise the release of the disclosure free output but the researcher has to prepare a good research question. We make the connexion because we have a 10-years' experience, establishing agreements with national statistical agencies. We also have our own staff who can evaluate the proposal and support the analysis with specific coding of the data and therefore coordinating this work. We construct a code to make a dataset for France, for example, making sure that this coding is useful for downloading data for The Netherlands. We are trying to make the link providing expertise facilitating the use of the data.

It works like this. The researcher comes up with a proposal and goes to CompNet MDI, which will engage with statistical institutes and prepare the micro data and probably the code to access the data and after the dataset is ready, prepare the code to utilise this data. We already have six countries in our project, In France we have direct access via your CASD and it is the same for the Netherlands. We use remote execution for Denmark, Finland, Norway, and Sweden. Eventually, the products of this coding will be aggregated by sectors but maintaining all the richness of the granular data at firm level. This will be key for researchers for their own analyses and publications.

As an example of what we have already done, we have been doing research that was proposed two or three years ago. We have business registers and Sylvain mentioned this codification. We can distinguish firms by their own identifier and connect these identifiers with a number of statistics that are already available and constructed by all statistical institutes in Europe. These are international trade statistics, international sourcing survey, foreign affiliate statistics, structural business statistics, community innovation surveys and ICT surveys. We created with the centre of business register identifiers and a metadataset that could be useful to dig deeper into the question of understanding productivity, analysis and connecting this with innovation and globalisation. This was the purpose of this project. We have used this metadataset for six countries, including France, and we now have data from 2010 to 2020 and we are expanding on a constant basis this data. This dataset is the starting point.

At this point, let me mention that we have an instrument, MDI, which can be extended to other countries and of course, to other values, such as education, etc., through the identifiers, as Sylvain said. We used this on Dutch data to look into a couple of research questions on what the general trends were in markups in the difference between the revenues and the costs. As we know, we want to have markups as small as possible in a competitive economy. What are the actual developments here? Can we explain differences in markups using different technologies and how qre they correlated with markups. Finally, are markups justified by higher productivity? Is it true that when there are higher markups thereis also higher productivity? I will just mention some results. Here, you can see markups data from the Netherlands from two datasets with very similar results. On the red line there is the revenue and on the blue line total costs. You can see that they are very much parallel, so markups in this dataset will be shown to be stable over time but only on average. This is the beauty of having granular data. If we actually look deeper at the level of categories of firms, the situation is much more heterogeneous than the other 2%. For example, what we have done is to separate manufacturers on the left-hand side from services on the right-hand side, and we split the firms depending on whether they are medium high-tech or low-tech. The results are pretty striking and clear, and we were expecting this. You can see that the markups of high-tech industries are higher than those of low-tech firms. This is the first result we got. It is true that on average markups are constant, but if you look at this by category of firms, you can see that in fact low-tech firms tend to be low markups and high-tech firms have higher markups. Certainly, as we would expect, Google, etc., make higher markups and it is the same thing for services.

What is the correlation of this sort of markups with the total factor productivity, which is an index of efficiency? Can we say that the higher the markups are the higher the productivity is? It looks likeit is not the case. If you look at the left-hand side, the quintiles of the markups, you can see there is actually a negative correlation. Higher productivity firms are not necessarily the ones extracting the higher markups. Therefore, having higher markups is not necessarily a sign of higher productivity. We are extending this to France and other countries because we want to inform competition policies and figure out whether these competition policies are sufficient to limit a new market power and foster productivity of the economy as a whole and how it changes from one country to another in Europe.

The conclusion of this is that there is something you know as statisticians: clear, granular datasets are essential to tackle issues such as the market power and productivity prospect. However, data availability is tremendously constrained across countries and this Compnet MDI project is aimed at creating robust and comparable datasets across Europe and facilitating their use. We count on the French statisticians, economists and experts, to keep being at the forefront. It is true that INSEE and the CASD are in fact template that we should follow. So we really hope that the example of France in both quality and access to the data will actually be spread to the whole of Europe. I count on your collaboration to make this happen.

I look forward to your reactions to the presentation and to your emails if you would like to know more.

1. L’utilisation des panels en statistiques d’entreprises

Sylvain MOREAU

Le dispositif SINE a pour objectif d’analyser les conditions de développement et les difficultés rencontrées par les entreprises nouvellement créées au cours de leurs 5 premières années d’existence. Créée en 2014, cette enquête a connu plusieurs éditions qui ont permis de bien analyser la façon dont le régime d’auto-entrepreneur a pris de l’ampleur, d’avoir une bonne appréhension du profil de ces créateurs d’entreprises et d’identifier le taux de survie des entreprises. Les questions s’enrichissent au fil des éditions. Pour 2022, des questions ont ainsi été ajoutées sur le traitement des déchets.

Autre exemple, le panel de groupes constitue un travail beaucoup plus exploratoire pour essayer de suivre de façon longitudinale des unités qui changent de périmètre. Nous publions tous les ans, au niveau sectoriel, des indicateurs agrégés qui ne permettent pas d'avoir de distribution. Pour la première fois l'an dernier, nous avons complété cette étude globale par une étude plus longitudinale sur 5 ans, pour voir comment la santé financière de l'entreprise avait évolué au niveau individuel. Ce panel de groupes permet de prendre en compte les évolutions des filiales. L'exercice soulève un certain nombre de questions sur la continuité économique. Les règles ont fait l'objet d'un débat au sein de l'Insee. Dans l'*Insee Références* sur les entreprises, présenté voilà un an, nous voyions bien que les secteurs qui restaient les plus fragiles étaient ceux qui allaient être le plus affectés par la crise Covid, dont l'hôtellerie-restauration. Cet outil est encore en construction aujourd'hui. Il fait l'objet de nombreux échanges pour rechercher les meilleures règles de continuité.

Il me semble important de montrer que ces outils sont très riches et qu'ils présentent un certain nombre de difficultés d'utilisation.

Je me pose une question : comment assurer une comparabilité entre les pays ? Il faut un travail spécifique sur l'unité légale. La situation est différente dans chaque pays.

Filippo di MAURO

As you say, defining the firm is obviously very important. What we have in our micro data infrastructure is central data available either by firm or by group. This [inaudible] that survey data are [inaudible], so they are published and [inaudible] bigger units. Typically, also balance sheet data are at the group level, so enterprise, [inaudible], etc. In general, we link the two levels in the system register and store the information on the group, the breakdown in different firms when needed. We try to have as much data as possible to match the firm in different destinations to the groups. Obviously, as you just said, each country requires ad hoc solutions, so we engage with statistical institutes to understand the best way going forward. For example, in the Netherlands we matched together two types of these [inaudible] registers, so-called NVR, the [inaudible], to obtain the right matching between firms and group. In fact, we are collaborating with the INSEE in [inaudible] where this deals with the data and actually understanding the best way of going about the problem. You point to this as a critical issue and we are doing this on a country-to-country basis using information we get from our national counterparts.

Nasser MANSOURI-GUILANI, CGT

Je suis économiste, ancien membre du Cnis, mais je m'exprime en mon nom personnel. Si j'ai bien compris, CompNet est un organisme privé. Au sein des statistiques publiques, n'avons-nous pas les moyens de conduire ce genre de travaux de façon plus pertinente, et moins coûteuse ? Ne voyons-nous pas là une sorte de privatisation des statistiques publiques ?

Filippo DI MAURO

C'est une bonne question, mais ce n'est pas le cas. CompNet est né dans la banque centrale européenne, avec un réseau de chercheurs il y a 6-7 ans. Nous sommes adossés à un institut économique allemand et nous avons un petit budget de la Commission européenne, la Banque centrale européenne, la Banque européenne d'investissement et d'autres organisations européennes intéressées par une meilleure structure des données au niveau européen. Nous travaillons tous pour le bien commun. La démarche ne répond pas à des intérêts privés.

Sylvain MOREAU

Au niveau national, la mise en qualité des données est faite pour la France. Nous faisons quand même un gros travail. Toutes les données sont mises à disposition au niveau individuel auprès du CASD. CompNet permet des comparaisons européennes. Qu'il existe un endroit où se trouve rassemblée la totalité de la documentation sur les données, permettant ainsi aux chercheurs d'appréhender la comparabilité, est essentiel. Certaines études sont totalement biaisées dès lors que l'unité de base est différente.

Une difficulté se pose sur la trace française des entreprises. Dans certains secteurs, cette trace est importante, mais si l'on veut aborder les questions de chaînes de valeur, il faut s'intéresser à un champ plus large. Je pense que ce travail ne peut pas être mené par l'Insee tout seul.

Nasser MANSOURI-GUILANI

Pourquoi ne travaillez-vous pas plus avec Eurostat ?

Sylvain MOREAU

Nous travaillons avec Eurostat. Nous pouvons rencontrer des difficultés liées au fait que les problématiques que nous avons en France ne sont pas les mêmes dans les autres pays. Ce travail pour la qualité des données est long, la France s'occupe des données françaises. A côté, un travail est fait au niveau européen par ceux qui peuvent nous aider.

Filippo DI MAURO

Une initiative lancée voilà quelques années, Amadeus Orbis, est une initiative privée qui met à la disposition des chercheurs des données collectées dans tous les pays. C'est une organisation privée qui n'a pas la capacité d'homogénéiser, de créer des bases de données au niveau national qui soient vraiment comparables. Il faut avoir une vision plus européenne pour les chercheurs qui sont intéressés par les comparaisons internationales et qui n'ont à la disposition que les données privées alors que nous sommes en mesure de générer des données bien meilleures. Il nous faut des données européennes pour assurer la comparabilité. Nous travaillons tous pour le bien commun européen.

Louis-André VALLET, CNRS et Sorbonne Université

Dans d'autres domaines, des initiatives ont été lancées par le monde universitaire pour fédérer des données statistiques nationales, notamment la Luxembourg Income Study et la Luxembourg Wealth Study. Ces données sont mises à la disposition de la recherche après avoir été regroupées et harmonisées. C'est donc là un autre exemple de collaboration fructueuse entre le monde de la statistique publique et le monde de la recherche universitaire.

IV. SESSION THEMATIQUE 3 : EDUCATION ET FORMATION

Thierry ROCHER, Service statistique du ministère de l'Education nationale et de la Jeunesse

Merci pour l'invitation à ce colloque. Je voudrais excuser Fabienne Rosenwald, directrice de la Depp. Nous proposons d'organiser cette session en deux séquences. Je vous présenterai les travaux de la Depp sur le sujet et nous entendrons deux points de vue. José Puig interviendra en tant qu'utilisateurs et nous aurons le point de vue d'un chercheur, Marc Gurgand, qui exposera son point de vue sur l'intérêt de ces données à travers quelques exemples concrets.

1. Panels et cohortes statistiques dans le domaine de l'éducation

La Depp a une histoire assez ancienne sur les panels sur échantillon et plus récente sur les cohortes de générations entières d'élèves. Les panels de la Depp existent depuis 50 ans, avec des points de départ en CP ou 6ème. Pour chaque panel, nous suivons plusieurs milliers d'élèves. Le dernier panel est entré en 2011 en CP.

Le handipanel suit les élèves en situation de handicap et a démarré en 2013-2014. Enfin, nous venons de lancer un panel d'élèves qui démarre en petite section de maternelle. Il s'agit d'un panel très prospectif avec l'accent mis sur les premiers apprentissages. Les panels de la DEPP présentent la particularité d'être très riches. Les enquêtes auprès des familles sont très précises sur les conditions matérielles, les attitudes et les pratiques des parents. Nous avons également ajouté des tests standardisés et plus récemment des enquêtes sur les pratiques pédagogiques des enseignants de maternelle. Les données recueillies par ces dispositifs de panels sont très utilisées par la recherche. Elles servent aussi à l'éclairage des politiques publiques.

Concernant le suivi de cohortes exhaustives, nous avons connu une histoire un peu mouvementée sur la question de l'identification des élèves. La situation s'est depuis stabilisée avec la mise en place de l'INE, l'identifiant national unique, et des systèmes d'information qui se sont considérablement améliorés et enrichis, avec un champ plus large et une meilleure qualité. Ce dispositif présente de nombreuses

possibilités d'exploitation. En particulier, les suivis de cohortes exhaustives nous permettent de produire des statistiques à un niveau plus fin, territorial ou même celui des établissements scolaires eux-mêmes.

En guise d'illustration, les IVAL, les indicateurs de valeur ajoutée des lycées, existent depuis 30 ans. Il s'agit de documenter les réussites des lycées et de les relativiser en fonction du contexte. Grâce aux suivis de cohortes exhaustives, nous avons pu enrichir les IVAL avec des données individuelles concernant le niveau initial des élèves à l'entrée au lycée (résultats au brevet national).

Plus récemment, nous avons travaillé avec la Dares sur la rencontre de deux cohortes : la cohorte des élèves et celle de l'emploi afin de voir ce que sont devenus les élèves grâce à la DSN. Les indicateurs sont parus récemment. Nous pouvons produire des indicateurs par établissement, par formation, avec le taux de poursuite d'études, le taux d'emploi à 6 mois, et en tenant compte du contexte et du profil des jeunes. Ce dispositif extrêmement riche offre des perspectives très importantes pour l'avenir.

Nous avons aussi utilisé deux dispositifs de panels afin d'évaluer les politiques publiques, dans une approche quasiment expérimentale. Nous pouvons suivre les progrès d'un échantillon d'élèves bénéficiaires d'un dispositif (réduction de la taille des classes de CP/CE1 par exemple) et les rapprocher des progrès d'un échantillon témoin, apparié sur le premier.

Enfin, nous avons pu estimer les effets de la crise sanitaire par la mise en place d'évaluations nationales pour documenter l'impact de la fermeture des écoles en comparant les cohortes avant, pendant et après. En complémentarité, nous avons lancé une étude de panel sur 1 000 écoles pour étudier les effets en essayant d'avoir une vision à 360 degrés (élèves, familles, enseignants, écoles).

Les panels sont beaucoup plus riches en termes d'information tandis que les cohortes statistiques proposent une déclinaison territoriale beaucoup plus fine. Cette masse de données soulève des enjeux d'accès aux données que ce soit en interne au ministère, en open data comme Inserjeunes et les IVAL, ou pour les chercheurs. Sur ce point, la DEPP est très active dans le projet IDEE, un équipement pour la recherche (Equipex) qui permet notamment de faciliter l'accès à nos données pour réaliser des projets de recherche. Nous espérons aboutir dans les prochains mois.

2. Panels et cohortes pour éclairer l'action publique en matière de scolarisation des jeunes en situation de handicap

José PUIG, Délégation interministérielle pour l'autisme et les troubles du neurodéveloppement

La stratégie nationale pour l'autisme est une politique publique pilotée par une déléguée interministérielle rattachée au premier ministre. Elle ne se limite pas, comme traditionnellement, à cantonner les questions de handicap à une question de santé. D'autres ministères sont impliqués : éducation, intérieur, justice, agriculture avec une gouvernance associant des représentants des ministères, des associations. Cette stratégie fait suite à trois plans autisme, chacun ayant fait avancer les réflexions sans pour autant régler les problèmes. Nous travaillons actuellement à la suite, une forme de 5ème plan. Nous ne parlons plus seulement d'autisme, mais de troubles neurodéveloppementaux. Cette politique a un coût de l'ordre de 450 millions d'euros sur 5 ans, plus un reliquat des crédits non utilisés dans le 3ème plan. Nous avons donc l'obligation de mesurer l'impact des 101 mesures de cette stratégie, en comprendre les effets à court et long terme et détecter les angles morts. Chaque nouvelle solution conduit à poser un nouveau problème. L'intérêt des indicateurs statistiques permet de sortir de deux logiques : une logique de communication politique pour mener une approche plus qualitative, avec une vue sur les conséquences de la scolarisation sur l'enfant et sa famille. Dans le domaine de la santé, il est également très important de veiller à prévenir toute forme de conflit d'intérêts.

La scolarisation relève de deux systèmes, le système scolaire et le système médico-social qui regroupe l'ensemble des établissements sociaux et médicosociaux. Ces deux mondes sont très séparés, voire opposés dans l'histoire. Ils dépendent de ministères avec des cultures différentes : une culture hiérarchique, avec une mécanique de déconcentration, alors que le système des établissements médicosociaux rassemble des opérateurs majoritairement de droit privé et associatifs qui travaillent suivant une logique d'agrément. Lorsque l'on demande aux personnes de travailler ensemble, cela peut entraîner des répercussions. Ce n'est que depuis 2005 que le rapprochement entre ces deux systèmes est encouragé.

En 1999, un plan Handiscol avait pris différentes mesures, notamment le rapprochement des outils statistiques des deux ministères, la Drees et la Depp avaient été invitées à réfléchir à la mise en place de

systèmes d'observation conjoints. Cela a abouti à un groupe de travail qui a tourné court. La loi de 2005 a remis les choses à plat, avançant dans la direction de cette meilleure collaboration entre le monde scolaire et le monde médicosocial sans reprendre comme un volet spécifique ce rapprochement d'outils.

Les études se heurtent à deux obstacles principaux. Le premier tient à la non-concordance des nomenclatures. Les catégories de personnes sont peu standardisées, instables et évolutives. Les catégories des dispositifs (classes ou établissements) peuvent être croisées pour fabriquer une typologie totalement artificielle des personnes selon les établissements censés pouvoir les accueillir, dérive dénoncée par les usagers et leurs associations. Il existe des classifications internationales, notamment la classification internationale des maladies de l'OMS, des terminologies propres à l'éducation nationale et d'autres aux ministères sociaux. Ces classifications évoluent dans le temps. Ainsi, la classification internationale arrive à sa 11ème version. Les administrations répugnent à modifier leurs classifications. Dans chaque administration, ces classifications ont des fonctions différentes (répartition des crédits de l'assurance maladie via les ARS). Cette classification résiste à des évolutions légitimées par la progression de la connaissance scientifique car elle vient contrecarrer des dispositions réglementaires. On conserve des classifications relatives à des agréments d'établissements. Les structures sont assez conservatrices.

L'utilisation de nomenclatures non concordantes crée un bruit dans la façon de regarder comment les choses évoluent. A cela s'ajoute un second obstacle. Il faut veiller à ce que les données respectent un certain nombre de conditions scientifiques, éthiques et techniques. Il faut utiliser des méthodes validées, des échantillons pour tester les questionnaires, mener une réflexion méthodologique avec des spécialistes, faire des tests de comparabilité entre les groupes pour voir s'il y a lieu de conduire des études comparatives avec quelles précautions, combiner études qualitatives et quantitatives. Enfin, il faut veiller au respect de la loi et s'assurer que les données vont être exploitées dans des objectifs conformes à ceux qui ont présidé à leur recueil. Les méta-analyses se développent beaucoup, pour procéder à des évaluations sur de très grands nombres.

Le handi-panel a démontré, en cherchant à voir le devenir des enfants scolarisés à l'école, que leur origine sociale pesait plus lourd que leur situation de handicap. Il faut donc conduire une politique inclusive, mais si l'on s'attaque aux inégalités sociales, on corrige aussi des inégalités pour des élèves en situation de handicap dont le handicap ne fait qu'accroître les écarts dus à leur origine sociale. Le panel permet par exemple d'apprécier l'accès au lycée.

Pour les unités UEMA, la Depp a lancé un petit panel pour voir si ces dispositifs très coûteux ont une efficacité correspondant aux attentes. Dans le cadre de la stratégie pour l'autisme, nous venons de lancer une cohorte de santé publique sur 2 300 familles d'enfants repérés comme présentant des risques de troubles neurodéveloppementaux. Cette cohorte permettra une amélioration de la connaissance. Nous lançons aussi une recherche sur les dispositifs d'autorégulation.

Ces dispositifs se sont multipliés. Il est important d'avoir des évaluations comparatives pour déterminer lesquels méritent d'être conservés ou supprimés. Cette évaluation longitudinale est néanmoins encore peu développée.

Marc GURGAND, Paris School of Economics

Les données administratives présentent une richesse incroyable sur les élèves et les établissements dans lesquels ils se trouvent. L'INE permet de panéliser très facilement l'ensemble de ces données et de suivre un grand nombre d'élèves sur des durées longues. Ces données sont une source absolument incroyable.

Il faut penser à trois usages. Le premier consiste à exploiter ces données en tant que telles. Nous pouvons suivre les élèves dans le supérieur. De nombreuses recherches ont été menées avec ces données, notamment sur l'orientation. Ces sources contiennent des informations qui peuvent être collectées à faible coût marginal (notes au brevet, évaluations, etc.). Ces informations sont disponibles presque sans attrition. Être capable d'enrichir des données d'enquête avec des données administratives est extrêmement précieux. Le 3ème champ, qui est mon champ de recherche, est l'expérimentation sociale. Au début, il fallait conduire des enquêtes sur des centaines ou quelques milliers d'élèves avec des tests en fin d'année. De plus en plus, les décideurs publics que ces évaluations intéressent nous demandent de tirer des enseignements sur le long terme. Or il faut être capable de suivre les individus dans la durée. Les effectifs étant généralement petits, l'attrition nous obsède. Faire du suivi de long terme à partir d'enquêtes est difficile. Le programme Progressa au Mexique, expérimenté dans les années 1990, y est parvenu.

Nous avons mené une expérimentation voilà dix ans sur l'internat d'excellence, un objet assez nouveau dans l'éducation nationale. Nous avons lancé une expérimentation, tirant au sort l'ordre dans lequel ont été admis les candidats, ce qui nous a permis d'obtenir un échantillon témoin. Au bout d'un et deux ans, nous avons mené des enquêtes auprès de ces élèves pour mesurer leurs compétences en mathématiques et en français. Or l'exercice est compliqué. Sur 400 élèves, la moitié était toujours scolarisée à Sourdun, mais les autres devaient être retrouvés. Les résultats à court terme étaient nuancés. Avec les données administratives de la Depp, nous avons apparié notre enquête de 400 jeunes avec les données exhaustives de la Depp. Nous avons utilisé les données du SIES après le baccalauréat puis Inserjeunes pour savoir s'ils avaient trouvé un emploi.

Nous avons découvert avec étonnement que les effets de l'internat d'excellence sur l'accès au baccalauréat sont spectaculaires. Dans l'échantillon qui nous intéresse, la proportion qui obtient un bac général est de 47 % contre 68 % pour les jeunes passés par l'internat. Nous continuons de les suivre et il semble que l'histoire se prolonge dans le supérieur. Ils se retrouvent beaucoup plus dans les filières sélectives, et encore plus pour les jeunes issus de l'immigration.

On a du mal à démontrer que l'éducation prioritaire a des effets remarquables. Or cet objet a des effets incroyables. Certes, il est très particulier, et sans doute pas reproductible, mais il démontre que cela est possible. Pour avoir un discours sur des destins, pouvoir combiner nos données d'expérimentation avec les données administratives de la Depp est très précieux. L'Equipex Idée a pour vocation notamment de favoriser ce genre de travail sur les données pour aller le plus loin possible dans l'évaluation des politiques publiques.

Mohamed Mouloud HADDAK

Jusqu'à présent, les politiques pour lutter contre les inégalités ont échoué en France. Dans l'évolution des indicateurs, que ce soit dans l'enseignement secondaire, au bac ou dans l'accès aux filières sélectives, d'autres dispositifs sont-ils à l'étude ?

Thierry ROCHER

J'ai évoqué la réduction des tailles en CP. Nous avons mis en place cette étude dès septembre 2017, quelques mois après l'annonce de cette mesure. In fine, les résultats ont pu apparaître un peu décevants, mais ils sont positifs. L'écart entre REP+ et autres classes s'est réduit. Ces élèves vont entrer en 6ème cette année. Nous en retrouverons donc potentiellement avec le résultat aux évaluations nationales, avec des perspectives d'études pour voir des effets sur le long terme.

Marc GURGAND

La question est extrêmement vaste et peut nous emmener extrêmement loin.

Grégoire REY

Est-il envisageable de mettre en place un protocole un peu générique ou automatisé avec des données scolaires pour enrichir ces cohortes ?

Thierry ROCHER

Oui, c'est pour cela que nous mettons en place le projet Idée. Sur le projet Elfe, des échanges sont en cours pour faire des tests d'appariement. Elfe ne comprend pas l'INE. Il faut donc travailler autrement. Pour Marianne, nous n'avons pas démarré. C'est le but de cette plateforme d'imaginer les connexions possibles.

Alain CAILLAUD

Pour quelles raisons l'INE est-il différent du numéro Insee ?

Thierry ROCHER

Je l'ignore. Il existait d'autres identifiants avant l'INE.

Clotilde LIXI, SIES

De mémoire, la séparation a été demandée par la CNIL. Dans le passé (jusqu'aux années 80), le numéro de l'élève était le n° de sécurité sociale (numéro Insee).

Mouloud HADDAK

Il me semble que pour les enfants, c'est le numéro des parents.

Thierry ROCHER

Non

Nicolas PROKOVAS

Avons-nous une idée du coût que représente l'éducation prioritaire par rapport à l'éducation normale ?

Marc GURGAND

Un rapport de la Cour des comptes a été établi. L'internat d'excellence double le coût.

V. TABLE RONDE : LES PANELS DANS LE DOMAINE DE LA SANTE ET LEUR GOUVERNANCE

Benoît OURLIAC, Service statistique du ministère des Solidarités et de la Santé (Drees)

L'état de santé évolue tout au long de la vie. Nous sommes exposés aux risques, avec des effets qui varient selon la durée. La dimension temporelle du suivi des individus est importante pourproduire de la connaissance d'intérêt général et éclairer la décision publique. Les panels, par construction, mobilisent des concepts et définitions communs pour faciliter l'appropriation des résultats par les usagers et favoriser la comparaison. Cela peut jouer sur la façon dont sont construits les panels.

Cette table ronde vise à présenter ces deux approches et discuter de la façon dont elles peuvent s'enrichir et se coordonner. Pour ce faire, nous avons réuni Marie Zins, responsable de la cohorte Constances, Grégoire Rey, de France Cohortes et Mireille Elbaum qui a co-écrit un rapport voilà 3 ans sur les cohortes en santé.

Avant leur intervention, je vous présenterai un petit état des lieux. La constitution de panels est une opération extrêmement lourde, surtout dans ce domaine où les données recueillies sont sensibles. Il n'existe pas de panel purement enquête à la Drees. Les deux principaux dispositifs d'observation sont constitués d'une seule vague d'interrogation. Il n'y a pas de suivi longitudinal des personnes. Toutes sont des enquêtes ponctuelles. Il existe une exception notable, l'enquête Santé et climat professionnel réalisée par la Drees et la Dares en 2006 et 2010, avec un suivi longitudinal somme toute limité.

En parallèle, la France dispose aussi d'un vaste entrepôt de données médico-administratives, le système national des données de santé (SNDS) qui rassemble toutes les consommations de soins sur une profondeur temporelle pouvant aller jusqu'à 20 ans. Il peut être considéré comme un panel, avec toutefois certaines limites. La consommation de soins est un proxy pas toujours satisfaisant de l'état de santé des personnes. En outre, elle occulte les situations de non-recours et ne renseigne pas sur les facteurs de risque. Enfin, le SNDS contient très peu d'informations sur la localisation précise des personnes.

L'enquête Santé 2019 doit être appariée avec le SNDS pour compléter le suivi longitudinal avec les consommations et devrait permettre d'alléger la charge d'enquête pour la prochaine édition. Autre projet emblématique, l'EDP Santé, constitué par un appariement entre l'EDP et le SNDS, qui ouvre des perspectives très intéressantes avec un échantillon extrêmement large et communique des informations très riches sur les trajectoires des personnes.

L'enquête EPICOV a été lancée dès le début de la crise d'abord par l'Inserm. Le Cnis a réagi très rapidement pour pouvoir monter cette enquête en un temps record. Elle a donné lieu à trois vagues et une 4e et dernière vague sera sur le terrain en septembre prochain. L'enquête a pu être réorientée quand la santé mentale est devenue un enjeu majeur pour mesurer la persistance des syndromes dépressifs, l'évolution des attitudes des populations vis-à-vis des gestes barrières et de la vaccination. Outre ce dispositif classique qui repose sur une enquête statistique, la Drees a participé activement à l'exploitation des données des systèmes d'information sur la crise, avec des estimations en temps quasi continu de l'évolution de l'efficacité vaccinale.

Toujours dans le domaine de la santé, un dispositif réalisé par la Drees depuis une quinzaine d'années était en cours au moment de la crise et a pu être réorienté pour poser des questions sur les attitudes des médecins généralistes vis-à-vis de cette crise, des mesures gouvernementales et leurs pratiques de consultation. 5 vagues d'interrogation ont eu lieu pendant la crise, donnant lieu à une quinzaine de publications.

Marie ZINS, Université Paris Descartes

Je suis très heureuse de pouvoir vous présenter Constances, qui n'est pas qu'une cohorte mais plutôt une Infrastructure de recherche centrée sur une large cohorte. Au sein de cette infrastructure nous avons mis en place un réseau d'experts pour donner du contenu scientifique à cette cohorte et comme Constances est une base très complexe, nous avons mis en place une équipe pour accompagner les chercheurs dans son utilisation.

Cette cohorte généraliste vise à travailler sur la santé en général : toutes les pathologies et tous les facteurs de risque avec un nombre suffisant. Elle a pour objectif de devenir une plateforme polyvalente de qualité pour travailler dans différents domaines de la recherche biomédicale et la santé publique : sur la surveillance épidémiologique, l'histoire naturelle des maladies, les facteurs de risques, les effets indésirables à long terme des médicaments. Aujourd'hui, elle comprend un échantillon de 220 000 personnes. Pour l'inclusion et le suivi, nous nous sommes appuyés sur le réseau des centres d'examens de santé de la sécurité sociale. En effet la CNAM gère une centaine de centres dans tout le pays et réalisent des examens de prévention. Sont éligibles à Constances les personnes qui vivent dans l'un des départements Constances, âgées entre 18 et 69 ans, tirées au sort entre 2012 et 2020 et affiliées à la Cnam.

Les volontaires sont inclus dans un centre de santé, avec un bilan de santé et de nombreux questionnaires, des bilans biologiques. Le suivi est assuré par un auto-questionnaire annuel avec des questions répétées d'une année sur l'autre, et une partie répétée de façon régulière (alimentation, activité physique, santé mentale) et nous avons mis en place un double appariement avec le SNDS d'une part pour tous les volontaires depuis l'existence même du SNDS et les bases de l'assurance vieillesse d'autre part qui permettent de calculer les pensions de retraite. Les données seront beaucoup plus précises sur les trajectoires et leurs revenus. Tous les 4 ans, nous réinvitons les volontaires pour de nouveaux examens.

Quand nous avons présenté Constances, les ministères ont demandé que l'on puisse dire des choses sur la santé des Français. Nous avons donc tiré au sort des volontaires, selon un sondage à probabilités inégales pour favoriser l'inclusion des personnes qui participent moins à ces enquêtes. Et pour mieux travailler sur les effets de sélection, nous avons tiré au sort une cohorte de référence pour laquelle nous avons des données appariées aux mêmes baes administratives que les volontaires actifs et avec la même antériorité.

Nous recueillons un très grand nombre de données, notamment des données très précises. Nous avons demandé aux volontaires de reporter tous leurs épisodes professionnels de plus de six mois, des données sociales sur le volontaire et son conjoint, des données de l'examen de santé, mesures de pression artérielle, poids et taille, audition, vision, souffle et, pour les 45 ans et plus, des neuropsychologues ont proposé des tests cognitifs et des tests physiques (vitesse de marche, équilibre). En fait notre travail consiste à enrichir constamment cette base de données, par exemple pour recueillir les historiques résidentiels des volontaires. Avec les adresses géocodées, nous pouvons faire des évaluations d'exposition environnementales, notamment à la pollution atmosphérique. Un autre exemple d'enrichissement de la base : nous avons codé 198 000 calendriers professionnels en nomenclatures françaises et internationales. Cela nous permettra d'apparier avec des bases des matrices emploi-exposition comme celles développées par Santé Publique France.

Les bases administratives présentent un coût d'accès important, car ces bases n'ont pas été conçues pour cela. Nous venons de finir de programmer 54 algorithmes de détection de cas, ces algorithmes sont au catalogue des données Constances. Nous travaillons à l'actualisation des données de carrière et le RGCU (répertoire de gestion des carrières unique) et nous avons pour objectif à moyen terme d'apparier la cohorte avec les bases fiscales. Nous avons également répondu à un Equipex pour génotyper toutes les données d'ADN de la cohorte.

Nous ne travaillons pas seuls. Comme je l'ai dit en introduction, nous bénéficions d'experts réunis dans des groupes pluridisciplinaires thématiques qui nous aident à avancer dans cet enrichissement permanent de la cohorte.

En ce qui concerne la sécurité de la base, la cohorte appariée avec le SNDS est mise à disposition au niveau du CASD. Nous donnons accès à des données brutes et à des données nettoyées. Nous venons de finaliser une application d'interrogation de la base.

Qui sont les volontaires ? 54 % des volontaires sont des femmes. 13 % de jeunes de 18 à 29 ans, la moitié étudiants entre 20 % et 22% dans les autres tranches d'âge. Sur les niveaux d'études, 25 % des volontaires sont sans diplômes.

Constances s'est mobilisée sur la recherche sur la pandémie dans le cadre du programme SAPRIS SERO. Dès le 1er avril 2020, nous avons envoyé des questionnaires aux volontaires sur internet. Nous en sommes au 5e questionnaire et un nouveau est prévu en septembre. Nous avons réalisé des sérologies répétées. Cela a donné lieu à des publications de modélisation en collaboration avec l'Institut Pasteur.

Constances a déjà été utilisé dans une dizaine d'équipe de recherche impliquée sur la pandémie. L'avantage est que nous avons toutes les données déjà collectées et que nous continuerons à suivre ces volontaires. La pandémie peut être vue comme une rupture de trajectoire.

140 projets ont été évalués positivement par notre conseil scientifique avec l'aval de notre comité de pilotage. La Cnav et la Cnam ont un droit de veto sur l'utilisation de leur base. Nous faisons partie de consortiums européens et internationaux. Nous avons des partenariats institutionnels, notamment avec la MILDECA, Santé publique France, la Drees notamment sur les délais d'attente, des prévalences pour Eurostat, etc. Nous participons également à la recherche clinique au travers de RHU ou d'IHU.

Mireille ELBAUM, Inspection générale des affaires sociales

Je vais vous parler d'une mission menée sur les cohortes épidémiologiques, conjointement par l'Igas et l'Inspection générale de l'éducation, du sport et de la recherche. L'un des constats qu'elle a réalisés à l'époque était une faible communication entre les sphères statistiques et des cohortes épidémiologiques. Compte tenu des fonds importants engagés en 2010 au titre du programme des investissements d'avenir (PIA), nous étions interrogés sur la viabilité à terme de ces projets. La mission n'a pas voulu se limiter à cet aspect, s'interrogeant aussi sur la place des cohortes dans le domaine de la santé.

Quelques sujets de différences d'approches méthodologiques pour commencer. Les cohortes sont un instrument emblématique et un élément d'identité professionnelle des épidémiologistes. Dans le monde de la recherche, avoir son outil, sa source de publication, sa source de partenariat est très important. Les agences sanitaires ont chacune un groupe de cohortes qu'elles financent et suivent. La validité recherchée est le suivi de l'état de santé des personnes de la façon la plus fine possible, pour des individus qui sont toujours les mêmes, et dont on connaît des prélèvements biologiques, le génome, le microbiote, etc. Les statisticiens voient des problèmes de représentativité liés au volontariat et à l'attrition lorsque ces cohortes sont fermées. Sous le dénominatif « cohortes » dans le monde de la recherche en santé existent des projets extrêmement différents les uns des autres. Le portail épidémiologique qui y est dévolu en a recensé plus de 250, sans être exhaustif. Ces opérations n'ont pas grand-chose à voir les unes avec les autres. Nous avons appelé à une clarification entre les grandes infrastructures de recherche, des cohortes spécifiques sur des patients diagnostiqués ou traités pour une pathologie précise (cancer de la vessie, greffe de moelle osseuse) et des« plateformes » de service constituées par exemple par les chercheurs sur les maladies rares.

Du point de vue des statisticiens, il existe de grandes différences selon que ce sont des cohortes ouvertes ou fermées, selon qu'elles soient rétrospectives, prospectives ou mixtes, qu'elles donnent lieu à des recueils de données actifs ou passifs. Nous avions constaté à l'époque que les sphères statistiques et les

chercheurs en épidémiologie communiquaient assez peu. Je suis ravie de voir que cela commence à changer. Nous avions déjà noté une exception très intéressante, le dispositif Elfe, une cohorte d'enfants avant la naissance. Elle a servi de plateforme puisqu'est venue s'y greffer une cohorte spécifique d'enfants prématurés. S'est ainsi créé un outil servant à plusieurs fonctions, qui pouvait accueillir tant des approches sociologiques sur les mécanismes de transmission des inégalités dans l'enfance que des préoccupations scientifiques et médicales extrêmement pointues.

Cette méconnaissance des deux sphères peut plus largement encore être questionnée aujourd'hui. Nous avons de plus en plus d'outils de suivi longitudinal individuel dans le domaine de la santé avec des outils pour partie partageables entre les deux sphères, médicale et statistique. Les registres présentent une dimension longitudinale ou servent d'appui à des cohortes, par exemple s'agissant des cancers pédiatriques. Les bases de données administratives, en particulier le SNDS, ouvrent des possibilités nouvelles. L'un des problèmes tient au fait que les données médicales n'y sont pas complètes. Il ne comprend pas encore, par exemple, le contenu des analyses biologiques. Ces informations permettraient, si elles étaient disponibles de réaliser des cohortes de grande taille, faire des échantillons témoins, effectuer des appariements qui peuvent aller jusqu'aux données socioéconomiques et géographiques très fines, à condition que l'accès en soit assuré sous l'égide au CASD qui accueille les données socio-fiscales.

Les règles d'accès au SNDS, bien qu'élargies, nécessitent la constitution de dossiers, le passage à la CNIL, la justification des utilisations demandées, et pour les appariements c'est encore plus complexe. Pour les grandes infrastructures en population générale, le jeu en vaut très largement la chandelle, mais la question peut se poser pour les petites cohortes. En outre, le RGPD a imposé de travailler dans des environnements complètement sécurisés.

A l'époque de la mission, nous avions vu émerger des projets concurrents de mutualisation et d'appui aux chercheurs, qui avaient avant tout mis l'accent sur l'accueil informatique des cohortes et de leurs appariements avec les données du SNDS, c'est-à-dire la constitution pour chacun d'un grand système informatique. Le *Health Data Hub* avait pour fonction de servir de sas entre le système, la recherche et les données de santé et d'aider les chercheurs notamment à réaliser des appariements. Ce projet a été freiné et n'est pas encore arrivé à maturité. L'Inserm a pris conscience de cela avec la création du projet France Cohortes. Le CASD avait une longueur d'avance sur le plan des équipements d'accueil, il présente l'avantage d'aider à réfléchir sur les appariements avec les données socio-fiscales et a accueilli Constances. Ce paysage complexe ne facilitait pas la tâche et l'orientation à privilégier pour les chercheurs.

La mission avait relevé des enjeux de mutualisation et d'amélioration de la gouvernance dans ce paysage dispersé, sachant que les porteurs de cohortes étaient aussi confrontés à des problèmes juridiques, de financement, de marchés, de conventionnement avec des partenaires privés extrêmement complexes, les porteurs de petits projets étant le plus désemparés face à ces enjeux. Nous avons souligné le besoin de priorisation dans les choix de financement public opérés en matière de cohortes épidémiologiques. Il ne s'agissait pas selon la mission de continuer à de lancer des appels d'offres *bottom-up* visant l'excellence du protocole de recherche sans priorisation sur les thématiques et les pathologies observées. Nous avons appelé à une cartographie de l'ensemble des outils disponibles en matière de données longitudinales en santé, qu'ils appartiennent au monde épidémiologique ou statistique. En termes de système cible, en comparant notamment avec le Royaume-Uni, nous avons considéré avoir au minimum besoin de deux grandes cohortes épidémiologiques appariées aux données du système statistique et du SNDS : une d'enfants dès ou avant la naissance et une autre d'adultes, avec une ouverture périodique de la cohorte ainsi constituée. La cohorte Constances n'a pas été ré-ouverte et ce n'est donc plus un projet qui entre dans le cadre du Cnis.

Nous avons également considéré que ces suivis longitudinaux ne devaient pas porter exclusivement sur l'état de santé, mais devaient être couplés aux parcours de soins. D'une part parce face aux inégalités socioéconomiques, les prises en charge ne sont pas les mêmes. En outre, certains sujets sont « à cheval » entre la santé et le social, comme la perte d'autonomie, le handicap. Nous avons appelé de nos vœux un cadre de pilotage partenarial s'appuyant sur une instance de coordination des dispositifs d'observation et de suivi longitudinal en santé et associant les organismes de recherche, les agences et le système statistique public.

Nous avions surtout demandé à ce qu'il y ait des éclaircissements sur le « qui fait quoi » entre le Health Data Hub, le CASD, France Cohortes, et à lever les freins à l'utilisation du SNDS et autres données de référence pour les outils déjà à l'œuvre, notamment pour des appariements. Et nous avons souhaité que des solutions soient apportées aux chercheurs qui se heurtaient à d'autres problèmes. L'appui juridique et

budgétaire, par exemple, mais aussi la structuration des bio-banques, avec des questions de stockage très importantes. Enfin, la doctrine comme les modalités concrètes doivent être précisées et affirmées, de concernant les partenariats noués avec des entreprises privées. Au cœur du PIA, existait en effet une incitation à nouer des partenariats public-privé. Ces partenariats ont eu lieu dans des conditions un peu différentes selon les cohortes, alors que les financements privés peuvent potentiellement influer sur les priorités retenues.

Grégoire REY, Institut national de la santé et de la recherche médicale

Je suis statisticien et épidémiologiste. J'ai été choisi comme directeur de France Cohortes. Je suis encore directeur du CépiDC pour quelques mois et à ce titre je suis un peu au fait des questions de statistiques publiques. Les relations avec la statistique publique me semblent indispensables.

Les cohortes permettent d'observer les populations longtemps, finement, en respectant les droits des personnes. A l'Inserm, le chercheur a la liberté de proposer son protocole sur une thématique donnée. Toute la difficulté pour l'administration de la recherche est de structurer ce qui peut l'être tout en permettant au chercheur de travailler son sujet. La législation des dix dernières années s'est renforcée et les moyens pour la suivre n'ont pas toujours suivi, faisant naître un besoin de structuration.

Il existe un grand nombre de cohortes en population générale et des cohortes cliniques qui suivent plutôt des patients. Pour se concentrer sur une maladie, la cohorte clinique est indispensable. Les deux exercices sont donc complémentaires. 1 % de la population française fait partie d'une cohorte Inserm. France Cohortes hérite des grands investissements d'avenir des années 2010 pour coordonner les grandes infrastructures de recherche et elle vise à trouver le dénominateur commun de ces projets pour aider à trouver la pérennisation des grandes infrastructures et consolider le niveau scientifique. Des cohortes sont associées à divers titres.

Les cohortes ne sont pas seulement des systèmes d'information. La sécurité reste néanmoins une exigence absolue. France Cohortes n'a pas vocation à refaire ce que d'autres font. Elle doit assurer l'expertise épidémiologique et statistique pour harmoniser et mutualiser les outils qui peuvent et doivent l'être. Au cours des dix dernières années, l'idée de produire de la donnée et d'en permettre la réutilisation pour d'autres s'est développée, à l'instar de ce que fait Constances. Beaucoup de cohortes reposent sur un projet porté par un chercheur avec un programme de recherche et un conseil scientifique, mais veillant à éviter toute redondance. France Cohortes ne va pas révolutionner la gouvernance des anciennes cohortes, mais va essayer de rendre visible cette gouvernance, les règles du jeu, l'accès aux données.

Les aspects réglementaires ne sont pas négligeables et nécessitent un support important, d'autant que les réglementations évoluent très régulièrement. France Cohortes est une infrastructure récente, tout comme le Health Data Hub. De fait, les structures se cherchent. Elles essaient d'éviter autant que possible la redondance par des échanges réguliers.

France Cohortes n'est pas une plateforme offrant un service de prise en charge totale des cohortes. Elle vient en support à la conception. Le responsable de la cohorte conserve la responsabilité de la conception et la réalisation d'un recueil clinique. Sur le recueil passif, les méthodes sont mutualisables. France Cohortes peut donc proposer de réaliser des traitements, les protocoles d'appariement, etc. Nous pouvons aussi être moteurs sur l'ouverture des données. Au-delà de rendre les données visibles, il faut communiquer avec ceux qui participent aux cohortes pour qu'ils aient le sentiment d'être impliqués et limiter l'attrition. Nous essayons de créer des groupes de travail thématiques avec des utilisateurs, pour partager les expériences, et aussi participer à l'amélioration des outils. Elfe est la 1re cohorte à entrer dans le système, c'est en cours. Nous prenons également en charge des bases (maladies rares) qui ressemblent plus à des registres que des cohortes ou encore la nouvelle Marianne portant sur la thématique de l'autisme.

S'agissant des relations avec la statistique publique, le décret 98-37 prévoit un protocole d'appariement autour de la recherche de statut vital au RNIPP et l'appariement des causes de décès. Or, malgré la création du Health Data Hub, la mise en œuvre du chaînage SNDS avec récupération du NIR est aujourd'hui bloquée depuis plusieurs années. Quand le NIR n'est pas recueilli à la source, il faut faire une requête. Un protocole porté par la statistique publique présenterait un intérêt.

Je pense que l'EDP Santé est une mine d'or sur de nombreux sujets. Il est important de clarifier les conditions d'accès à ces données pour que les chercheurs puissent y accéder sans avoir besoin d'une

procédure ubuesque. Pouvoir se comparer avec la population générale me semblerait intéressant, dans l'esprit de Constances.

Epicov constitue un excellent exemple. Cette cohorte a bénéficié de la collaboration avec le service statistique public. Dans ce cas, des portes s'ouvrent. Je plaide pour que l'on puisse étendre cette démarche et trouver un modus operandi sans attendre une nouvelle crise sanitaire.

En matière de documentation, il est important de s'inscrire dans un processus commun de description des données avec les standards de l'Insee. Sur l'échantillonnage, la sphère statistique est plus habituée et sur la méthodologie d'inférence causale nous avons nous aussi des choses à partager. Tout ceci forme un beau programme de travail.

Benoît OURLIAC

Sur la reconstitution du NIR, j'imagine que vous avez en tête la pétition lancée par des chercheurs. Dans le cadre de l'EDP Santé, nous partageons pleinement l'objectif de mettre ce panel à la disposition du monde de la recherche. La démarche a été ralentie un temps, la Cnil y voyant la constitution d'un nouvel entrepôt de données de santé. Des projets mobilisent déjà l'EDP Santé au sein de l'IrdesS, responsable de traitement. Nous avons communiqué auprès du monde de la recherche sur le mode d'emploi pour réaliser cet accès à l'EDP Santé. A moyen terme, nous souhaiterions le mettre à la disposition de la recherche comme d'autres enquêtes de la statistique publique.

Sur Epicov, ce qui a ouvert les portes c'est qu'il s'agit d'une enquête de la statistique publique et je pense que le Cnis accueillera volontiers toutes les enquêtes présentées en opportunité, mais cela implique quelques contraintes.

Mireille ELBAUM

L'Autorité de la statistique publique essaie de faire entrer dans la régulation d'ensemble de la statistique publique les producteurs de statistiques en santé. Nous avons fait entrer Santé publique France pour certaines séries statistiques en matière d'incidence et de prévalence d'un certain nombre de pathologies. Leurs enquêtes devraient passer au Cnis pour obtenir le label, dont ils ont pris conscience notamment en termes méthodologiques, car il y a des sujets qui peuvent donner matière à contestation.

Christel COLIN, Insee

En effet, cette question de la récupération du NIR pour d'autres objets a été identifiée, mais nous n'avons pas d'offre à proposer pour l'instant. Nous en avons pris conscience fortement au cours des derniers mois. Nous devons y réfléchir.

Mireille ELBAUM

La Cnam sert de porte d'entrée à côté du Health Data Hub. Sur les aspects informatiques et pour les appariements, seule la Cnam en est responsable et elle est embolisée, y compris pour ses partenaires les plus proches, même l'Irdes. Elle est restée pour le moment le guichet d'entrée unique.

Grégoire REY

Il existe un cas d'application sur lequel l'Insee peut agir : l'appariement du SNDS avec les causes de décès. Si je résume, France Cohortes devrait porter une action collective auprès du Cnis ?

Benoît OURLIAC

Toutes les enquêtes visant à recueillir des informations d'intérêt général peuvent avoir vocation à solliciter l'avis d'opportunité et le label du Cnis. C'est de ce point de vue un guichet ouvert.

Grégoire REY

J'imagine qu'il existe des exigences sur la gouvernance des données ?

Benoît OURLIAC

Sur la gouvernance en général, non, mais bien sur la méthode, la qualité et la diffusion des résultats, ainsi que la mise à disposition des données.

Mireille ELBAUM

Nous retombons sur une question centrale. Les sujets de représentativité et d'attrition font partie des critères examinés par le comité du label.

Grégoire REY

Je parlais de finalités différentes de celle de représenter la société française, comme faire de l'inférence causale : ce n'est pas éligible au Cnis ?

Mireille ELBAUM

Les cohortes de patients n'entrent pas forcément dans la statistique publique. Il existe un ensemble de cohortes représentatives en population générale avec une vocation intermédiaire, avec le sujet essentiel de la réouverture. Dès lors que l'on rouvre, la nature du financement devient le problème le plus difficile à résoudre car il faut mobiliser des financements et des partenariats en continu.

Marie ZINS

Encore une fois, toute cohorte n'a pas vocation à aller au Cnis. Nous avons toujours intérêt à savoir pourquoi une personne est entrée ou non dans une cohorte.

Patrice DURAN

Toutes les cohortes n'ont pas besoin de passer au Cnis. La statistique publique répond à des règles extrêmement précises. L'un des problèmes pour Epicov, c'est que tous les résultats sont publics.

Laurent TOULEMON

L'Ined ne fait pas partie de la statistique publique. On est un institut de recherche, nous passons quelques enquêtes au Cnis, mais nous lançons aussi une enquête européenne qui obéit aux contraintes de financements européens. Sur les délais, nous avons fixé une durée maximale de trois ans après la collecte. Nous ne sommes pas dans la même logique. Les chercheurs qui mènent des enquêtes doivent publier. L'Insee a un rôle central en termes d'aspect quantitatif et ce rôle est difficile à jouer, car il n'est pas dans ses missions aujourd'hui. Il y a un sujet de mutualisation d'outils, non dénués de concurrence parfois.

Mireille ELBAUM

C'est sur les infrastructures longues, mixtes, comme Elfe ou Nutrinet, dont on ne sait pas si elles seront rouvertes que la question se pose. Quand elles servent de base à certains travaux épidémiologiques, on recherche avant tout la causalité, mais ces infrastructures se situent dans un entre-deux.

Marie ZINS

Nutrinet n'a jamais voulu être représentatif, c'est un dispositif de recherche pour étudier l'effet de l'alimentation sur la santé. Il y a une relecture de l'histoire. Constances a été mis en place dès l'origine pour le partage des données, pas Nutrinet. Nous pouvons regretter l'absence de cohorte représentative en France sur ce sujet.

Grégoire REY

Je reviens sur l'exemple du décret 98-37 avec des chaînages en mode service. Nous n'avons pas forcément besoin de passer devant le Cnis pour faire un chaînage avec le SNDS ou récupérer des données fiscales. Il faudrait clarifier le sujet. Nous devons proposer une recette sans avoir à passer devant le Cnis. Je conçois que ce n'est pas simple.

Benoît OURLIAC

Quand les SSM souhaitent enrichir les enquêtes avec le code statistique non signifiant, ils présentent leur projet au Cnis.

Mireille ELBAUM

L'ASP a pris un délibéré sur le sujet des appariements internes au SSP. A priori, chaque appariement n'a pas vocation à suivre la même procédure que les enquêtes mais devra être indiqué dans le programmes annuel des organismes.

CLOTURE

Christel COLIN, Insee

Le directeur général participe à une réunion européenne avec ses homologues. Il me revient donc de conclure cette journée riche en présentations et en échanges. Je tiens à remercier tous les intervenants. Je ne vais pas tenter une conclusion très élaborée, mais une synthèse de points transversaux.

L'un des premiers objectifs du colloque était de montrer les avancées depuis le rapport Chaleix-Lollivier. Les exposés ont montré la richesse des sources développées récemment ou en construction et la diversité croissante des thématiques que les panels et cohortes permettent d'explorer. Les panels historiques (EDP, DADS, panels d'élèves de Education nationale) existent toujours et ne cessent de s'enrichir (périmètre, intégration de nouvelles sources, etc.). Ce développement permet de réaliser des analyses plus fines et de répondre à de nouvelles questions. D'autres panels se sont développés. Nous avons vu des exemples aujourd'hui avec des enquêtes répétées dans le temps, éventuellement complétées par des données administratives pour combler les trous, et des panels administratifs, pour du suivi passif, complété dans certains cas par des enquêtes directes. C'est le cas du panel des élèves de la Depp, de l'Eniacrams pour les bénéficiaires de minima sociaux. Cette diversité des panels tient aussi à la diversité des modes de construction.

J'ai également été frappée par l'immense diversité des observations de tranches de vie. Il faut observer longtemps pour suivre les trajectoires, mais nous avons aussi des exemples de panels courts intéressants, comme SRCV qui dure 4 ans. L'enquête Emploi avec ses 6 interrogations trimestrielles est aussi un panel. Inversement, les panels tous salariés permettent par exemple de suivre sur des dizaines d'années des personnes ayant commencé leur carrière en 1976.

Les thématiques sont de plus en plus diversifiées. Dans certains domaines comme la santé, les outils sont foisonnants. Ils sont également nombreux sur les thèmes de l'emploi ou de la formation. Mais on peut noter avec beaucoup d'intérêt des projets nouveaux sur des champs jusqu'alors peu couverts : du côté de la Justice et de l'Intérieur comme vu ce matin, du côté des entreprises avec le développement de suivis longitudinaux.

Ces avancées visent toutes à répondre à des questions, sociales ou économiques. Il est important de bien identifier la question à laquelle on veut répondre. Les panels et cohortes sont indispensables pour répondre à certains besoins de connaissance ou pour l'évaluation des politiques publiques. Ils présentent aussi l'intérêt de croiser les thèmes, avec des panels de plus en plus multithématiques pour aborder différentes dimensions de la vie des personnes. Mais la question de l'objectif, des finalités, doit rester essentielle dans le développement de ces outils. Car en regard, se posent des questions de protection de la vie privée. Avec les panels on accumule de l'information, parfois très riche ou très personnelle, sur certaines personnes, parfois nombreuses. Le Cnis a toujours été attentif à soutenir le développement d'outils d'observation longitudinaux tout en veillant au respect de la confidentialité et de la protection de la vie privée. On en a parlé lors de la rencontre du Cnis de janvier dernier sur les appariements. Les potentialités d'appariement ne cessent de se développer, avec des inquiétudes de la société civile. Cela renforce aussi l'exigence de transparence. Par le passé, beaucoup de panels étaient sur échantillon pour éviter de construire de larges entrepôts.

Les panels sont indispensables pour éclairer certaines questions, mais ce sont des dispositifs compliqués et qui exigent des moyens. Ils posent la question des identifiants. En outre, la multiplication des sources intégrées dans les panels multiplie aussi les problèmes potentiels d'appariement : plus on ajoute des

sources, plus le risque d'avoir des défauts d'appariement augmente. Au-delà des questions juridiques, il faut donc suivre toute une méthodologie. L'appariement n'est pas une opération magique. Même avec un Siren unique, le suivi des entreprises ne va pas de soi : une entreprise peut se restructurer, intégrer un groupe, se séparer de certaines activités, etc.

Du côté des panels par enquête, il faut pouvoir suivre les individus, y compris quand ils déménagent, les convaincre de répondre plusieurs fois. Dans la statistique publique, nous n'avons pas recours en général aux incitations financières. Il faut traiter l'attrition par des méthodes appropriées, rafraîchir les échantillons pour qu'il restent représentatifs. L'exploitation est également complexe. Tout cela a un coût. Il ne faut pas l'oublier. Ce coût peut être d'autant plus un sujet de difficultés que les panels requièrent un investissement sur le temps long. La pérennité en dépend et nous connaissons tous des panels arrêtés pour des problèmes de moyens.

Dernier point, la question de la coordination des acteurs : les panels demandent souvent une coordination des différents acteurs à la fois pour la mise en place des cohortes et panels, parfois en interministériel (Intérieur et Justice, par exemple, ou encore Inserjeunes, qui va s'étendre au supérieur…). Nous avons vu que les cohortes en santé nécessitent souvent la mise en place et la structuration de communautés de chercheurs. Cette coordination est également nécessaire pour éviter le foisonnement et garantir une certaine efficience, créer des synergies.

J'ai bien noté la liste de courses adressée à la statistique publique. La statistique publique a évidemment un rôle à jouer sur différents aspects : la mise en place, l'entretien, le développement de panels, leur exploitation, les développements méthodologiques, la mise à disposition des données pour les chercheurs, la mise en place de groupes d'exploitation. Il est aussi frappant de voir la place importante du monde de la recherche dans ces panels, et ce, dans différents domaines. Cela amène souvent la recherche et la statistique publique à collaborer et forcément on pourrait faire plus. Le monde de la recherche tient aussi une place particulière dans les cohortes internationales, comme Share ou le programme Générations et Genre.

Nous venons de mettre en ligne une étude de l'Insee sur la mobilité intergénérationnelle des revenus à partir de l'échantillon démographique permanent, qui met en regard la position dans l'échelle de revenus des jeunes adultes avec celle qu'occupaient leurs parents.

Je tiens à remercier vivement l'ensemble des intervenants et participants, ainsi que les initiateurs et organisateurs de ce colloque. C'est grâce à eux que ce colloque, d'abord programmé en 2020, a pu se tenir aujourd'hui.

La séance est levée à 17 heures 05.